Optimal group sequential tests

Eales, John D.

Award date:
1991

Awarding institution:
University of Bath

Optimal Group Sequential Tests

submitted by J.D. Eales
for the degree of PhD
of the University of Bath
1991

John O. Eales

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract.

Sequential procedures were originally designed for use in an industrial context. However the flexibility and efficiency of sequential methods made them attractive to those involved in medical experimentation.

The earliest sequential designs for clinical trials were fully sequential, that is they required an analysis to be conducted after every patient response. More recently the emphasis has been on group sequential designs, where analyses are carried out after groups of patient responses.

One of the distinguishing features of sequential procedures is that the required sample size is a random variable. For fixed group sizes, a given maximum number of analyses and given error constraints, group sequential tests can be designed which minimize a given function of expected sample size. We term such procedures **optimal group sequential tests.**

In this thesis we introduce a computationally efficient and numerically stable method for the computation of optimal group sequential tests. Although we approach this problem from a frequentist perspective, our method makes use of both Bayesian decision theory and dynamic programming.

In Chapters 3 and 4 we consider computing optimal one-sided and two-sided tests respectively. The two-sided tests permit the rejection of the null hypothesis, H_0, at any analysis, but they only allow H_0 to be accepted at the final analysis. In Chapter 5 we consider computing optimal wedge tests which, like two-sided tests, test H_0 against a two-sided alternative, but, unlike two-sided tests, allow H_0 to be accepted or rejected at each analysis.

In Chapter 6 we consider some of the Bayesian and Bayes decision theoretic procedures proposed in the literature. Finally, in Chapter 7, we look at a number of ideas for future research as well as some relevant topics which have not been considered elsewhere in the thesis.

Acknowledgements

This research was carried out under the supervision of Dr. Christopher Jennison . I would like to express my thanks to him for his help, encouragement and kindness during the course of this work.

Thanks are also due to my friend Julian Stander for all his help, including his comments on earlier versions of Chapters 3 and 4.

Other friends who deserve thanks for their encouragement and help include Mario, Howard, Wei, Katerina, Guy and Glenn.

My family has been a constant source of encouragement and support over the last 26 years. I would like to take this opportunity to express my sincere gratitude to them.

Finally it is my pleasure to acknowledge the Science and Engineering Research Council for their financial support over the years 1987-1990.

Dedication.

This thesis is dedicated to the memory of my grandmother, Hilda Jane Williams (July 21st 1910 - February 2nd 1989).

Table of Contents

1. Sequential Methods: An Introduction.

1.1 Introduction.

Johnson (1961) defined a sequential method to be '... any statistical procedure in which the final pattern (including the number) of observations is not determined a priori but depends, in some way or other, on the values observed in the course of the work'. Johnson's definition covers the theory, methods and examples of this thesis.

1.2 The History of Sequential Methods.

The first sequential test can be traced back to 1929 when a paper by Dodge & Romig entitled 'A Method of Sampling Inspection' was published in the Bell System Technical Journal. Dodge & Romig proposed a double sampling inspection procedure for use in acceptance sampling.

It was in April 1943 that Abraham Wald introduced the **sequential probability ratio test** (SPRT) (which we describe in detail in §3.4). The SPRT laid the foundations for many of the recent developments in the theory of sequential analysis. At about the same time as Wald, G.A. Barnard was carrying out war-time work in Britain of a sequential nature.

The main motivation for the work of Wald and Barnard came from industrial problems such as quality control. As an example, consider the following problem: a random sample of size n is drawn from a batch of manufactured goods. A rule is applied so that if r ($< n$) of the sample are in some sense 'defective' we reject the batch; otherwise the batch is accepted.

Clearly, in the majority of cases, it will prove quite unnecessary to inspect all n items in the sample before coming to a decision. Hence the total number of items inspected will not be fixed, but a random variable taking values in the range $[r,n]$.

This simple example highlights many of the properties of sequential methods:-

(i) the method is efficient in the sense that it employs as few observations as possible before coming to a decision;

(ii) the sample size is not fixed, but a random variable;

(iii) the method leads to savings both in terms of time and money over the corresponding fixed sample size procedure.

Sequential Methods in Clinical Trials.

More recently researchers have been interested in employing sequential methods in clinical trials. The nature of clinical trials is such that patients respond gradually over time. The investigator will naturally want to analyse the resulting data as it becomes available in order that the trial may be stopped as soon as an important difference between treatments becomes apparent. However the sequential nature of such an analysis must be borne in mind when deciding whether or not to stop the trial.

Armitage (1960) was one of the pioneers in the field of sequential clinical trials. Later, Armitage, McPherson & Lowe (1969) introduced the notion of repeated significance testing at a constant nominal significance level. These early procedures were fully sequential, that is they required an analysis to be carried out after every patient response. Such an approach is, of course, likely to prove impractical in the context of most clinical trials.

Pocock (1977) and O'Brien & Fleming (1979) were among the first to propose group sequential tests, where analyses are conducted after a group of patient responses. Both of these papers concentrated on the two-sided testing problem, where a null hypothesis, H_0, is tested against a two-sided alternative, H_1. Rejection of H_0 in favour of H_1 is permitted at any analysis by these designs, but acceptance of H_0 is only allowed at the final analysis. More recently Fleming, Harrington & O'Brien (1984) and Wang & Tsiatis (1987) have also proposed two-sided group sequential tests.

Gould & Pecore (1982), Gould (1983) and Emerson & Fleming (1989) have proposed group sequential wedge tests, where again we are testing H_0 against a two-sided alternative, but this time H_0 can be accepted or rejected at each analysis. DeMets & Ware (1980), (1982), Whitehead (1983), Jennison (1987) and Emerson & Fleming (1989) have proposed one-sided group sequential tests.

The big advantage of a group sequential test over a fixed sample size test lies in the fact that, on average, it will normally require fewer patients. This is particularly the case when treatment differences are large.

Recently there have been a number of papers in the literature relating to the computation of optimal group sequential tests. For a given problem with fixed group sizes, a fixed maximum number of groups and given error constraints, the optimal test minimizes a chosen function of expected sample size.

Jennison (1987) used a constrained numerical search method to compute optimal one-sided tests. Unfortunately his method suffered from a number of computational problems (see §3.5 for a description of these problems).

Wang & Tsiatis (1987) considered a family of two-sided tests indexed by a single parameter, Δ, say. They then searched over Δ for the optimal test within their family. Their approach was computationally efficient and gave near optimal tests for any given problem.

Emerson & Fleming (1989) used an analogous approach to that of Wang & Tsiatis in order to compute near optimal one-sided tests.

In this thesis we introduce an improved method for the computation of optimal group sequential tests. Unlike the approach of Jennison (1987), our method is computationally efficient and numerically stable. Further, unlike the approaches of Wang & Tsiatis (1987) and Emerson & Fleming (1989), our method gives the overall optimal test for a given problem.

Despite the obvious attractions of group sequential tests and, in particular, optimal group sequential tests, O'Brien & Fleming (1979) have noted a reluctance to use formal sequential methods on the part of clinicians. They gave 3 main reasons for this:

(i) the analysis of fixed sample size tests is well known and generally thought to be efficient;

(ii) the complex nature of some study protocols for sequential trials;

(iii) the experimenter may wish to stop the trial for reasons unrelated to treatment differences (for example, the development of a new, seemingly better, treatment).

A number of papers in the literature have proposed less rigid, more flexible, sequential methods for use in practice. For example, Lan & DeMets (1983), (1989) and Jennison (1987) have suggested methods for use when group sizes and/or the maximum number of groups are unpredictable. We also note that Pocock (1977) showed that his two-sided tests are robust to departures from the usual normality and known variance assumptions, as well as being robust to small

changes in the planned group sizes.

Jennison & Turnbull (1984), (1989) have proposed an extremely flexible procedure known as the repeated confidence interval (RCI) approach. The RCI approach allows a trial to be stopped early for reasons unrelated to treatment differences, while not invalidating any inferences which might be made concerning the size of treatment differences.

Bayesian and Bayes Decision Theoretic Sequential Methods.

The emphasis in this thesis is on frequentist sequential methods. A number of Bayesian sequential methods have been proposed in the literature by, for example, Berry (1987) and Freedman & Spiegelhalter (1989). Bayesian decision theory approaches have been suggested by, among others, Anscombe (1963), Chernoff & Petkau (1981) and Berry & Ho (1988).

1.3 Summary of the Thesis.

The examples and motivation for the work presented in this thesis come from the area of medical statistics. As an introduction to this topic we consider the design of clinical trials in Chapter 2.

Chapters 3, 4 and 5 deal with the one-sided, two-sided and wedge testing problems respectively, from a frequentist perspective. In each chapter we give a computationally efficient and numerically stable method for computing optimal group sequential tests.

Chapter 6 considers some of the Bayesian and Bayes decision theoretic procedures proposed in the literature. We look in detail at the frequentist properties of one of these procedures and compare them with the frequentist properties of some of our optimal tests.

Chapter 7 looks at a number of ideas for future research as well as some relevant topics which have not been considered elsewhere in the thesis.

2. Clinical Trials.

2.1 Introduction.

Pocock (1983) defined a clinical trial to be '... any form of planned experiment which involves patients to elucidate the most appropriate treatment of future patients with a given medical condition'. Much of this thesis will be concerned with comparing different designs for clinical trials. In this chapter we consider some of the more general aspects of clinical trial design.

2.2 Designing Clinical Trials.

An Example

Consider comparing the relative effectiveness of a new drug, N, say, with that of a standard, S, for reducing the blood pressure level in patients suffering from hypertension. Before such an experiment is conducted the new drug will have passed through a number of preliminary checks and pilot studies concerning its safety and efficacy. Initially these checks will be carried out in test tubes or on animals. Eventually, however, the drug will have to be tested on humans. Following Pocock (1983), this stage of the experiment, the clinical trial, can be divided into four distinct phases:-

Phase I: Clinical Pharmacology and Toxicity.

Here the main concern is with drug safety rather than drug efficacy. Phase I trials are normally conducted on healthy human volunteers. Typically these trials require between 20 and 80 volunteers, although separate studies in this phase may require as few as 6 volunteers.

Phase II: Initial Clinical Investigation for Treatment Effect.

Here relatively small scale investigations are conducted into both the efficacy and safety of a new drug. This phase is useful for two main purposes:
(i) to screen out any drugs which are seen to be either ineffective or over-toxic;
(ii) to obtain an optimum treatment policy in terms of doses and schedules.
Phase II trials rarely require more than 100-200 patients per drug.

Phase III: Full-Scale Evaluation of Treatments.

When people talk of clinical trials they are often referring to Phase III trials. Here a new drug which has passed through the first two phases of experimentation is compared with a standard drug, a placebo or another new drug. A "substantial number" of patients will be required by a typical Phase III trial.

Phase IV: Post-Marketing Surveillance.

After a new drug has been approved for marketing it will continue to be assessed in terms of both long-term morbidity and mortality effects. This is formally termed Phase IV of a clinical trial.

Primarily we shall be concerned with Phase III trials in this thesis.

The Legal Requirement for Clinical Trials.

Up until the 2nd World War there were no formal requirements for clinical trials before a drug could be freely marketed. However the thalidomide disaster in the early 1960s led to a tightening up of the legislation in both the U.S.A. and this country.

In the U.K. a Committee on Safety of Drugs was set up in 1963 with the aim of considering all new drugs before they were tested or marketed. However pharmaceutical companies were not legally bound to seek this Committee's approval. In 1968 the Medicines Act was passed. Part of this Act set up the Committee on Safety of Medicines (CSM). Any new drugs had now to be approved by the CSM before being included in clinical trials **and** before being placed on the market.

The Design of Clinical Trials.

Randomization.

On entry to a trial a patient is randomly assigned to one of the available treatments. There are two main reasons for randomization:

(i) to guard against bias;

(ii) to provide a basis for standard frequentist techniques such as significance tests.

A naive way of randomly allocating patients to treatments would be to toss a coin and to assign the patient to the new treatment if the coin falls "heads" and to the standard treatment if the coin falls "tails". Alternatively a table of random numbers could be used in place of the coin. Unfortunately such approaches do not guarantee equally sized treatment groups. Indeed, especially in the case of a small trial or of a group sequential test, serious imbalances might occur in the sizes of different treatment groups.

One way of ensuring treatment group balance is to use a matched pairs design. Here a pair of patients who are similar in terms of such characteristics as age, sex and social class, are entered on to a trial with one of the pair being randomly assigned to the new treatment and the other to the standard treatment. One logistical problem with a matched pairs study concerns the difficulty in pairing patients, particularly near the end of a study.

Restricted randomization offers an alternative way of achieving equally sized treatment groups. Whitehead (1983, Ch.2) suggested two methods of restricted randomization, **random permuted blocks** and **biased coin designs**. Both of these methods are described in detail by Whitehead.

Historical Controls.

An alternative to randomization would be to use **historical controls**. Here all the patients in a trial receive the new treatment and their responses are compared with the historical responses of patients who had received the standard drug before the start of the trial.

Whitehead (1983, Ch.2) has strongly attacked the use of historical controls. He argued that the historical data would have come from patients who had '... been treated during a different time period when the provision of secondary treatments, the standard of care and staffing, the administrative policy of the hospital or clinic and many other aspects of their welfare would have been different.' He went on to note that '... the record forms of the controls would not have been filled in for the purpose of the trial and might be incomplete, inaccurate or just inadequate for making the observations that can be recorded as they occur for the treatment group.'

There is much agreement with Whitehead. Available evidence suggests that inferences based on trials using historical controls tend to inflate the value of a

new treatment.

This does not mean that historical data is of no use. On the contrary this data can be used when planning a trial to estimate such design parameters as the variance of patient responses and the required sample size.

Within and Between Patient Studies.

In a within patient study each subject receives every one of the treatments being tested. This is achieved either by administering the treatments simultaneously or by administering them sequentially. For example, if we were comparing two treatments for earache we could simultaneously administer one treatment to a patient's left ear and the other treatment to a patient's right ear. Obviously logistics will often rule out a simultaneous within patient study. The alternative approach is to randomly assign a patient to one of the treatments. Then, after a suitable length of time, the patient is taken off this treatment and a "wash out" period follows to remove any lingering effects of the treatment. The patient is then given the other treatment (or randomly assigned to another treatment in cases where more than two treatments are to be compared). Such a design is known as a crossover trial.

The big advantage of a within patient study is that each patient acts as his or her own control. Also the theory behind crossover trials is relatively simple.

Disadvantages include the impossibility of within patient studies for certain trials (consider comparing a treatment involving surgery with one that does not, for example). Also crossover trials should be avoided if there is any possibility that treatment effects could be carried over from one course of treatment to the next.

In a between patient study each subject receives only one of the treatments being tested. Patient responses on the first treatment are then compared with patient responses on the second treatment and so on. Between patient studies are much more common than within patient studies. Gore (1982) noted that '... of 38 clinical trials reported in the *Lancet* over six months, 28 compared treatments between groups of patients'.

Blindness.

There are various degrees of blindness. In a single-blind trial the patient does not know which treatment he or she is receiving. In a double-blind trial both the patient and the doctors involved with the trial do not know which treatment the patient is being given. If at all possible a double-blind trial should be employed as knowledge of the treatment to be allocated might effect the patient psychologically and the doctors' decision to admit patients on to the study.

Fixed Sample Size and Group Sequential Designs.

Traditionally clinical trials have required a single sample of patients. Standard statistical techniques are then used to analyse the responses of these patients. The desire for a greater degree of flexibility in clinical trial designs has led to a number of group sequential procedures being proposed in the literature. Here an analysis of the accumulated data is carried out after a group of patients has responded. The trial is stopped if there is significant evidence to suggest that one of the treatments is superior to the other(s). Otherwise the next group of patients is admitted on to the trial. The trial is terminated at the Kth analysis.

Care has to be taken when designing group sequential tests in order to ensure that the Type I and Type II error rates are preserved. In the rest of this thesis we shall consider the design of group sequential tests for clinical trials. We shall pay particular attention to the design of optimal group sequential tests which minimize chosen functions of expected sample size over tests satisfying the required error rates.

3. Optimal One-Sided Group Sequential Tests.

3.1 Introduction.

In this chapter we consider one-sided hypothesis tests on the mean of a normal distribution with known variance. In §3.2 we consider the fixed sample size approach to this problem. In many areas of application, for example clinical trials, there exist strong arguments for employing a group sequential rather than a fixed sample size approach. A formal description of a one-sided group sequential test is given in §3.3, while the literature on one-sided sequential tests is reviewed in §3.4.

In §3.5 we introduce **optimal one-sided sequential tests** or, more simply, **optimal tests**. Jennison (1987) described a constrained numerical search method for computing these optimal tests. Unfortunately his method suffers from a number of computational problems. So as a compromise between computational considerations and optimality, Jennison (1987) proposed a second, computationally efficient, method for computing **near** optimal tests. Emerson & Fleming (1989) also proposed a computationally efficient method for computing near optimal tests.

In §§3.6–3.8 we describe an improved method for computing optimal one-sided group sequential tests. Our method has a number of important numerical and computational advantages over Jennison's original approach. Furthermore, unlike the second approach of Jennison and that of Emerson & Fleming, our method does give the actual optimal test for a given problem.

In §§3.9 and 3.10 we give some results for our method, while in §3.11 examples of optimal one-sided tests are considered. Practical problems associated with group sequential experiments are addressed in §§ 3.12 and 3.13.

Before proceeding we note that although the motivation for this work comes from a clinical trials problem, the implications and applications of optimal tests go much wider.

3.2 The Fixed Sample Size Test.

Consider a clinical trial designed to compare the relative efficacies of an experimental treatment (which we shall denote by E) and a control treatment or placebo (which we shall denote by C). The fixed sample size approach to this

problem requires a total of N_f' pairs of patients. On entry to the trial one patient in each pair is randomly assigned to treatment C with the other being assigned to treatment E. Let the random variable X_i $(i = 1, 2, \ldots, N_f')$ denote the difference in response between the ith patient on treatment E and the ith patient on treatment C, and suppose that the X_i's are independent and normally distributed with unknown mean μ and known variance σ^2. We wish to test

$$H_0: \mu \leq 0 \quad vs \quad H_1: \mu > 0$$

with error rates

$$\Pr(\mathcal{A}_1 \mid \mu = 0) = \alpha \tag{3.2.1}$$

and

$$\Pr(\mathcal{A}_0 \mid \mu = \delta) = \beta \tag{3.2.2}$$

where $\mathcal{A}_0$ and $\mathcal{A}_1$ denote the acceptance of hypotheses H_0 and H_1 respectively. The Type I error or size of the test, α, Type II error, β and δ - termed the reference improvement by Whitehead (1983, Ch.4) - are under the control of the experimenter. The number of pairs of patients required by the test, N_f', is then a function of α, β, σ^2 and δ and is given by equation (3.2.3)

$$N_f' = \frac{\sigma^2}{\delta^2} \{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2. \tag{3.2.3}$$

The test accepts H_0 if

$$S_{N_f'} = \sum_{j=1}^{N_f'} X_j \leq \sqrt{N_f' \sigma^2}\, \Phi^{-1}(1-\alpha)$$

and rejects H_0 in favour of H_1 otherwise.

In the context of a clinical trial there are a number of reasons for adopting a group sequential rather than a fixed sample size approach to the above hypothesis testing problem.

To begin with patients generally enter a trial sequentially over time and so patient responses tend to accrue gradually over several months or years. The natural curiosity of the experimenter often leads to a desire to conduct periodic analyses of the available data. The trial is stopped at any one of these analyses if

a "large" treatment difference is observed. Unfortunately, if the interim analyses are unplanned, the overall error rates of the test may be greatly increased. A properly designed group sequential test gives the experimenter the flexibility to carry out early analyses of the data, while at the same time controlling the probabilities of making errors.

Secondly, in the case of a "large" treatment difference, it is desirable to stop a trial at the earliest possible opportunity in order to minimize the number of patients who will receive the inferior treatment. Group sequential tests can be designed which, on average, allow for a study to be stopped earlier than a fixed sample size test with the same error rates. There is, therefore, an ethical argument for adopting a group sequential approach in clinical trials.

Thirdly, there is an important economic argument for employing a group sequential test. Clearly clinical trials are very expensive to run and in general it will be far more cost effective to offer a patient the standard treatment than to admit him or her on to a trial. An approach that attempts to minimize the number of patients entering an experiment will, therefore, be desirable from an economic perspective.

3.3 One-Sided Group Sequential Tests.

In this section we give a formal description of a one-sided group sequential test on the mean of a normal distribution with variance, σ^2, assumed known.

Consider again the one-sided hypothesis testing problem described at the start of §3.2. This time, however, a maximum of K groups of n pairs of patients are available for entry on to the trial with one patient in each pair being randomly assigned to treatment C and the other to treatment E.

Let $S_{in} = X_1 + X_2 + \ldots + X_{in}$ $(i = 1, 2, \ldots, K)$ denote the sum of responses from the first i groups. Clearly S_{in} is normally distributed with mean $in\mu$ and variance $in\sigma^2$. At analysis i $(i = 1, 2, \ldots, K)$ we decide on the basis of S_{in} whether to either stop the trial with the acceptance or rejection of H_0, or to continue entering patients on to the study. At analysis K the trial is terminated with H_0 either being accepted or rejected.

We shall consider stopping rules of the general form:-

At analysis i $(1 \le i \le K-1)$,

if $S_{in} \ge c_i$ stop entering patients on to the trial and accept H_1;

if $S_{in} \le c_i'$ stop entering patients on to the trial and accept H_0;

if $c_i' < S_{in} < c_i$ enter the next group of n pairs of patients on to the trial.

At analysis K,

if $S_{Kn} \ge c_K$ stop entering patients on to the trial and accept H_1;

if $S_{Kn} \le c_K$ stop entering patients on to the trial and accept H_0 .

The set of critical values or boundary points of the test, $\{(c_1',c_1), \ldots, (c_{K-1}',c_{K-1}), c_K\}$, is chosen to satisfy the error constraints (3.2.1) and (3.2.2). That is we determine critical values such that:

$$\sum_{j=1}^{K} \Pr\left(c_1' < S_n < c_1, \ldots, c_{j-1}' < S_{(j-1)n} < c_{j-1},\ S_{jn} \ge c_j \mid \mu = 0\right) = \alpha \quad (3.3.1)$$

$$\sum_{j=1}^{K} \Pr\left(c_1' < S_n < c_1, \ldots, c_{j-1}' < S_{(j-1)n} < c_{j-1},\ S_{jn} \le c_j' \mid \mu = \delta\right) = \beta. \quad (3.3.2)$$

The joint probabilities in the above equations can be expressed in terms of multiple integrals:

$$\Pr\left(c_1' < S_n < c_1, \ldots, c_{j-1}' < S_{(j-1)n} < c_{j-1}, S_{jn} \in r_j \mid \mu\right)$$

$$= \int_{r_j} \int_{c_{j-1}'}^{c_{j-1}} \cdots \int_{c_1'}^{c_1} f_\mu(x_1) \ldots f_\mu(x_j - x_{j-1})\, dx_1 \ldots dx_{j-1}\, dx_j$$

where $r_j = (-\infty, c_j']$ or $[c_j', \infty)$ and $f_\mu(x)$ is a normal density with mean $n\mu$ and variance $n\sigma^2$. These multiple integrals may then be evaluated numerically using an approach based on Simpson's rule (see Appendix 3.1 for details).

We shall term any set of critical values satisfying the error constraints (3.2.1) and (3.2.2) a **feasible** set and the corresponding test a **feasible test.** For $K \ge 2$ there are infinitely many feasible sets of critical values satisfying equations (3.2.1) and (3.2.2) for fixed α, β, δ and n.

Unlike fixed sample size tests the sample size of a sequential test is not determined *a priori* but is a random variable with a corresponding probability distribution. Of particular interest are the expected number of pairs of patients required by a test under some treatment difference μ (denoted by $E(N|\mu)$) and $E(N|\mu)$ averaged over some prior distribution for μ. For the stopping rule defined earlier $E(N|\mu)$ is given by equation (3.3.3)

$$E(N|\mu) =$$

$$n\sum_{j=1}^{K} j \Pr\left(c_1' < S_n < c_1, \ldots, c_{j-1}' < S_{(j-1)n} < c_{j-1}, S_{jn} \notin (c_j', c_j) \mid \mu\right). \quad (3.3.3)$$

The joint probability in equation (3.3.3) can be expressed as a sum of multiple integrals which can be evaluated numerically (see Appendix 3.1 for details).

A further important property of a sequential test is its maximum sample size. In our example a maximum of Kn pairs of patients are admitted on to the experiment. It is clear that Kn must exceed the corresponding fixed sample size, N_f', given by equation (3.2.3). Logistical considerations will often restrict our attention to group sequential tests with Kn only a few percent greater than N_f'.

For a given pair of error constraints (3.2.1) and (3.2.2), feasible tests may be compared in terms of their expected and maximum sample sizes. We shall consider such a comparison in later sections.

3.4 A Review of the Literature on One-Sided Sequential Tests.

Fully Sequential One-Sided Tests.

There have been a number of one-sided sequential tests proposed in the literature. The earliest was the sequential probability ratio test (SPRT) of Abraham Wald (1947). The SPRT was designed for use when testing the quality of batches of industrial components and it is fully sequential (i.e. a test is conducted after every observation). The boundaries of the test are given by

$$c_i' = \frac{i\delta}{2} + \frac{\sigma^2}{\delta}\ln A$$

and

$$c_i = \frac{i\delta}{2} + \frac{\sigma^2}{\delta}\ln B,$$

where the constants A and B $(0 < A < B)$ are chosen to give a feasible test. Wald showed that choosing

$$A = \frac{\beta}{(1-\alpha)} \quad \text{and} \quad B = \frac{(1-\beta)}{\alpha}$$

led to a test with error rates approximately equal to α at $\mu = 0$ and β at $\mu = \delta$.

As can be seen the continuation region of the SPRT is defined by a pair of parallel lines of slope $\delta/2$. Although there is no upper bound on the maximum sample size of the test, Wald showed that the SPRT will terminate after a finite number of observations with probability one. Wald & Wolfowitz (1948) proved that the SPRT is optimal in terms of minimizing both $E(N|\mu=0)$ and $E(N|\mu=\delta)$ over the set of feasible tests. Clearly, though, the infinite maximum sample size together with the requirement that data monitoring is fully sequential makes the SPRT most unsuitable for use in clinical trials.

Anderson (1960) noted that the SPRT is sub-optimal in terms of its expected sample size for μ in the range $(0,\delta)$. Indeed the SPRT may, on average, require more observations than the corresponding fixed sample size test here. Anderson considered designing fully sequential tests which, compared with the SPRT, reduced $E(N|\mu)$ for $\mu \in (0,\delta)$ while not leading to substantial increases in $E(N|\mu=0)$ and $E(N|\mu=\delta)$. He was particularly interested in tests which reduced $E(N|\mu=\delta/2)$ compared with the SPRT.

To this end he considered a family of tests with critical values of the form

$$c_i' = a_1 + b_1 i$$

and

$$c_i = a_2 + b_2 i$$

where $a_1 < 0 < a_2$. To ensure that the experiment stops after a finite number of observations, sampling is terminated after K observations with the acceptance of H_0 if $X_1 + \ldots + X_K \leq c$, a constant, and the rejection of H_0 in favour of H_1 if $X_1 + \ldots + X_K > c$. It would seem intuitively reasonable to have $b_2 < 0 < b_1$ in order that the lines defining the continuation region of the test are converging towards each other.

Of course, for fixed a_1, b_1, a_2, b_2, K and c, we could easily calculate the error rates and expected sample size for the Anderson test using the numerical

methods described in Appendix 3.1. Back in 1960, however, the calculation of error probabilities and expected sample sizes for sequential tests was a major computational problem. Anderson overcame this problem by considering an analogous problem in continuous time. He replaced the test statistic by a Wiener process, $X(t)$, which is a gaussian process such that $E(X(t)) = \mu t$, $Var(X(t)) = t$ and $Cov(X(t'),X(t'+t)) = t'$ for any t, $t' \geq 0$. The stopping boundaries for the continuous time problem are given by $a_1' + b_1' t$ and $a_2' + b_2' t$, with sampling being terminated at some time T.

Anderson gave formulae for calculating the error rates and the expected sample size of his continuous time problem. These formulae can be used to give good approximations to the error rates and expected sample size for the original problem with its integer group sizes.

The most important test considered by Anderson was the triangular test, which has critical values given by

$$c_i' = -\frac{2\sigma^2}{\delta} \ln\left[\frac{1}{\alpha+\beta}\right] + \frac{3\delta i}{4}$$

and

$$c_i = \frac{2\sigma^2}{\delta} \ln\left[\frac{1}{\alpha+\beta}\right] + \frac{\delta i}{4}.$$

The triangular test is very efficient in terms of its expected sample size under $\mu = \delta/2$.

Group Sequential One-Sided Tests.

More recently a number of one-sided group sequential tests with finite maximum sample sizes have been proposed in the literature.

DeMets & Ware (1980) proposed a group sequential version of the SPRT with a maximum of K groups of n pairs of patients. The critical values of their test for $i = 1, 2, \ldots, K-1$ are given by

$$c_i' = \frac{in\delta}{2} + \frac{\sigma^2}{\delta} \ln\left[\frac{\beta}{(1-\alpha)}\right]$$

and

$$c_i = \frac{in\delta}{2} + \frac{\sigma^2}{\delta} z_U$$

with

$$c_K' = c_K = \frac{Kn\delta}{2} + \frac{\sigma^2}{\delta} z_U .$$

They chose z_U to give a test of size α and the group size, n, to give a test with Type II error β at $\mu = \delta$. The resulting test leads to substantial savings in both $E(N|\mu=0)$ and $E(N|\mu=\delta)$ over the corresponding fixed sample size test.

DeMets & Ware (1980) proposed two further one-sided group sequential tests. The first of these is a one-sided version of two-sided group sequential test of Pocock (1977), which we describe in detail in §4.4. DeMets & Ware termed this test **the one-sided group sequential method.** The critical values of the test are of the form : $c_i = \sqrt{ni\sigma^2}\, z$ and $c_i' = -c_i$ $(i < K)$, with $c_K' = c_K = \sqrt{Kn\sigma^2}\, z$. Again z is chosen to give a test of size α and n to give a test with Type II error β at $\mu = \delta$.

Clearly the one-sided group sequential method has only a small probability of stopping early under $\mu = 0$. However savings in $E(N|\mu=\delta)$ over the corresponding fixed sample size test are quite substantial.

The asymmetry inherent in the one-sided problem led DeMets & Ware to consider a third test with an asymmetric stopping rule. Somewhat predictably they termed this test **the asymmetric group sequential method.** DeMets & Ware pointed out that in most clinical trials less evidence is required to stop sampling and accept the null hypothesis (i.e. accept that the control treatment is no worse than the experimental treatment) than to reject it in favour of the alternative hypothesis. With this in mind they recommended choosing $c_i = \sqrt{in\sigma^2}\, z_U$ and $c_i' = -\sqrt{in\sigma^2}\, z_L$, $(i = 1, 2, \ldots, K-1)$ with $c_K' = c_K = \sqrt{Kn\sigma^2}\, z_U$. Here z_L $(< z_U)$ is free to be chosen by the experimenter and z_U $(> z_L)$ is subsequently chosen to give a test of size α. Again the group size, n, is chosen to give a test with Type II error β at $\mu = \delta$.

For fixed K, α and β, the expected sample size under $\mu = \delta$ for this test is similar to that for the one-sided group sequential method. However $E(N|\mu=0)$ is in general smaller for the asymmetric test.

In comparing their 3 tests DeMets & Ware noted that the group sequential version of the SPRT requires, on average, fewer **analyses** under $\mu = 0$ and $\mu = \delta$ than the other two tests. However it also requires larger group sizes in order to satisfy the Type II error constraint. Overall the group sequential version of the

SPRT was viewed as the best of the three tests because it led to large savings in $E(N|\mu=0)$ and small savings in $E(N|\mu=\delta)$ compared with the other two tests.

In a later paper DeMets & Ware (1982) proposed a test with the same lower boundary as the group sequential version of the SPRT and the same upper boundary as the one-sided version of the O'Brien & Fleming (1979) test (the two-sided O'Brien and Fleming test is described in §4.4). The critical values of this test are given by $c_i = z_{OBF}$ and

$$c_i' = \frac{in\delta}{2} + \frac{\sigma^2}{\delta}\ln\left[\frac{\beta}{(1-\alpha)}\right]$$

for $(i = 1, 2, \ldots, K-1)$ and $c_K' = c_K = z_{OBF}$. The constant z_{OBF} is chosen to give a test of size α, while the group size, n, is chosen to give a test with Type II error β at $\mu = \delta$. The resulting test has a relatively large probability of stopping early to accept H_0. However the probability of the test stopping early to reject H_0 in favour of H_1 is small.

DeMets & Ware (1982) claimed that their test was in tune with the views of many clinicians. Trials involving unpromising experimental treatments are stopped as quickly as possible in order to minimize the use of experimental resources. On the other hand, trials involving promising experimental treatments are allowed to continue so as to enable secondary issues, such as treatment side effects, to be assessed.

Whitehead & Stratton (1983) (also Whitehead (1983, Ch.4)) have described a group sequential version of the triangular test of Anderson (1960). In defining their stopping rule, Whitehead & Stratton made use of a continuity correction due to Siegmund (1979) and Cuzick (1981). The critical values of the group sequential triangular test are given by

$$c_i' = \frac{3\delta in}{4} - \frac{2\sigma^2}{\delta}\ln\left[\frac{1}{\alpha+\beta}\right] + 0.583\sqrt{n}$$

and

$$c_i = \frac{\delta in}{4} + \frac{2\sigma^2}{\delta}\ln\left[\frac{1}{\alpha+\beta}\right] - 0.583\sqrt{n}.$$

The test is very efficient in terms of its expected sample size under $\mu = \delta/2$.

Clearly there is a good deal of interest in the literature in designing feasible sequential tests which are efficient in terms of the expected number of patients entering a trial under some given treatment difference, μ. The advent of modern computers allows us to consider tests which minimize given functions of expected sample size over feasible stopping rules. We shall consider the computation of such tests in §3.5.

3.5 Optimal One-Sided Sequential Tests.

For a given problem with n, K, α, β and δ fixed, we are often interested in computing the sequential test which minimizes some given function of expected sample size known as an objective function. We shall term such a test an **optimal sequential test**, or, more simply, an **optimal** test. Jennison (1987) pointed out that the computation of stopping rules for optimal tests is eased if we choose the Type II error of our test, β, equal to the Type I error, α. Jennison also concentrated on the following symmetric testing problem:

$$H_1^-: \mu < 0 \quad vs \quad H_1^+: \mu > 0$$

with error rates

$$\Pr(\mathcal{A}_1^- \mid \mu = \delta) = \alpha \tag{3.5.1}$$

and

$$\Pr(\mathcal{A}_1^+ \mid \mu = -\delta) = \alpha \tag{3.5.2}$$

where $\mathcal{A}_1^-$ and $\mathcal{A}_1^+$ denote the acceptance of hypotheses H_1^- and H_1^+ respectively. Stopping rules for this problem are of the general form:

At analysis i $(1 \le i \le K-1)$,

if $S_{in} \ge c_i$ stop entering patients on to the trial and accept H_1^+;

if $S_{in} \le -c_i$ stop entering patients on to the trial and accept H_1^-;

if $-c_i < S_{in} < c_i$ enter the next group of n pairs of patients on to the trial.

At analysis K,

if $S_{Kn} > 0$ stop entering patients on to the trial and accept H_1^+;

if $S_{Kn} < 0$ stop entering patients on to the trial and accept H_1^- .

There is no major loss in generality from considering symmetric tests. For fixed K and α, the group size, n, and critical values of the optimal test $\{c_1, c_2, \ldots, c_K\}$, are easily rescaled to give an optimal test for problems with different values of δ and σ^2 (details of the necessary rescalings are given in Appendix 3.2).

The fixed sample size test for the symmetric problem requires N_f pairs of patients, where N_f is a function of α, δ and σ^2 and is given by equation (3.5.3)

$$N_f = \frac{\sigma^2}{\delta^2} \{\Phi^{-1}(1-\alpha)\}^2. \qquad (3.5.3)$$

The test accepts H_1^+ if $S_{N_f} > 0$, while it accepts H_1^- if $S_{N_f} < 0$.

Optimal Fully Sequential Tests.

Lai (1973) was one of the first to consider the computation of optimal tests. He considered minimizing $E(N|\mu=0)$ over feasible fully sequential tests with no bound on the maximum sample size. Weiss (1962) showed that a feasible test, $\mathcal{T}$, minimizes $E(N|\mu=0)$ if, for some $p \in (0,1)$, $\mathcal{T}$ minimizes

$$\Psi(\mathcal{T},p) = \frac{p}{2} \Pr(\mathcal{A}_1^+ \mid \mu=-\delta) + (1-p)\, E(N \mid \mu=0) + \frac{p}{2} \Pr(\mathcal{A}_1^- \mid \mu=\delta).$$

Lai used the result of Weiss to obtain a system of equations which can be solved numerically to give the desired optimal test. Clearly this approach could be extended in order to minimize other objective functions. Lai did not give any results for his method. Indeed Lorden (1976) in discussing Lai's approach noted that '... heavy computational work is required to carry out this algorithm, rendering it unsuitable for routine use'.

Lorden did, however, use Lai's algorithm to obtain results with which his fully sequential 2-SPRT test could be compared. The 2-SPRT stops sampling after i observations and accepts H_1^- if

$$S_i = X_1 + X_2 + \ldots + X_i < \frac{\sigma^2}{\delta} \ln A + \frac{i\delta}{2}.$$

Similarly it stops sampling after i observations and accepts H_1^+ if

$$S_i > -\frac{\sigma^2}{\delta} \ln A - \frac{i\delta}{2}.$$

To obtain error rates close to the desired α at $\mu = \pm\delta$, Lorden suggested choosing

$$A = \frac{\alpha}{0.4996 - 0.28645\delta + 0.0696\delta^2}$$

He reported that for α between 0.001 and 0.1 the attained error rates based on this formula for A are accurate to within 1/5000 for δ between 0.1 and 0.5 with a slight loss in accuracy as δ increases beyond 0.5.

The 2-SPRT is close to optimal in terms of minimizing $E(N|\mu=0)$. To demonstrate this point Lorden considered a range of values of δ and α and noted that the test was always within 1% of the overall minimum of this objective function.

Optimal Group Sequential Tests.

Jennison (1987) considered computing optimal group sequential tests for the following four objective functions:

$$F_1\colon E(N|\mu=0)$$

$$F_2\colon E(N|\mu=\delta)$$

$$F_3\colon E(N|\mu=2\delta)$$

and

$$F_4\colon \frac{1}{5}\{E(N|\mu=0)+E(N|\mu=\delta/2)+E(N|\mu=\delta)+E(N|\mu=3\delta/2)+E(N|\mu=2\delta)\}.$$

Clearly, for stopping rules symmetric about zero, objective function F_2 is identical to $E(N|\mu=-\delta)$, F_3 is identical to $E(N|\mu=-2\delta)$, and so on.

For a given objective function, F, and for fixed K, n, α, δ and σ^2, Jennison employed a constrained numerical search method to compute his optimal tests. (We note here that Jennison considered the rather special set of problems with $\sigma^2 = 1$ and $\delta = \Phi^{-1}(1-\alpha)$ so that $N_f = 1$. He went on to consider an analogous problem in continuous time. However the critical values of his stopping rules are easily rescaled to give stopping rules for tests with other values of δ and σ^2.) The search was conducted over the first $K-2$ critical values with c_{K-1} being constrained at each stage of the search to give a feasible stopping rule. A simple bisection search gave the relevant c_{K-1}. In cases where no such c_{K-1} existed, F

was assigned a "large" positive value in order to move the search away from infeasible regions.

The search algorithm employed was dependent upon the dimensionality of the minimization problem. When $K = 2$, assuming a sensibly formulated problem, there is a unique test satisfying the group size and error constraints. When $K = 3$ there is a single unconstrained critical value, c_1, and the minimization problem is one-dimensional. The Golden Section Search algorithm was used in this case. When $K > 3$ the minimization problem is multi-dimensional and the simplex algorithm of Nelder & Mead (1965) was used.

Unfortunately Jennison's approach is both computationally expensive and numerically unstable. For instance with $K = 10$ it can take up to 12 hours on a Sun-4 to converge to an optimal test. Further the Nelder & Mead algorithm is not particularly powerful in more than about 7 dimensions (i.e. $K = 9$) and does not guarantee convergence to the global minimum of F.

To overcome these problems Jennison considered the **error spending functions** of his optimal tests. An error spending function is simply the rate at which a test spends its Type I error expressed in a functional form. For his optimal tests Jennison noted that the error spending functions are sigmoid shaped and therefore are of the general parametric form

$$f(i;\alpha,\underline{b}) = \begin{cases} \alpha\left[1+\exp\left\{-\left(-\frac{b_1}{in}+\frac{b_2}{(K-i)n}+b_3 in+b_4\right)\right\}\right]^{-1} & 1 \le i < K \\ \alpha & i = K. \end{cases}$$

Here $\underline{b} = (b_1, b_2, b_3, b_4)$, b_1 and b_2 are constrained to be positive and $f(i;\alpha,\underline{b})$ gives the total Type I error spent by analysis i. Given $f(i;\alpha,\underline{b})$ $(i = 1, 2, \ldots, K-1)$ the corresponding set of critical values are computed by numerical integration and the bisection method. Again c_K equals zero.

For a given problem an objective function, F, can be minimized over the family of error spending functions of the above parametric form. The requirement that $f(K;\alpha,\underline{b}) = \alpha$ reduces the number of free parameters in this minimization problem from 4 to 3. Therefore the Nelder & Mead simplex algorithm is used to search over $(\ln b_1, \ln b_2, b_3) \in \mathbf{R}^3$ while b_4 is constrained at each stage of the search to give a feasible test.

Jennison reported that the test minimizing F over this family of error spending functions is very close to the overall optimal test. Also, for fixed K, α and F, varying the group size, n, leads to only small changes in the values of b_1, b_2 and b_3 for the optimal error spending function (although b_4 might vary quite considerably). Jennison tabulated the values of b_1, b_2 and b_3 leading to optimal error spending functions for each of the objective functions F_1, F_2, F_3 and F_4, for $\alpha = 0.01$ and 0.05, $K = 5$ and 10, and with group size n_{opt} (the group size which minimizes F when everything else is fixed). This table can be used to give near optimal stopping rules for other group sizes and values of K. For the examples considered by Jennison, attained minima were never more than 0.7% of the fixed sample size from the overall minima.

Emerson & Fleming (1989) considered minimizing objective functions F_1 and F_2 over stopping rules with critical values of the general form

$$c_i = (in)^{\Delta} z - in\delta \quad (i = 1, 2, \ldots, K).$$

Given K, α and δ, Emerson & Fleming searched over Δ for the minimum of the chosen objective function. At each stage in the search z was constrained to give a feasible stopping rule and n was constrained so that $c_K = 0$.

Clearly the Emerson & Fleming approach is more computationally efficient than the original approach of Jennison. For $K \leq 10$ and $\alpha = 0.01$ or 0.05 Emerson & Fleming showed that their tests were approximately 1% of the fixed sample size from the overall minima for F_2 and approximately 0.5% of the fixed sample size from the overall minima for F_1. The Emerson & Fleming approach can be seen to be a one-sided version of that of Wang & Tsiatis (1987) described in §4.5.

In §3.6 we propose a new method for the computation of optimal one-sided group sequential tests. Our new method is an improvement upon that of Jennison (1987) in terms of both computational efficiency and numerical stability.

Our approach considers a family of group sequential problems in Bayesian decision theory with a common prior distribution and cost of sampling function. The forms of the prior and cost function are determined by the objective function we are interested in minimizing. Individual problems within the family differ in their loss function. These loss functions are indexed by a single parameter, d. We show that by searching over d we obtain a loss function which gives a Bayes

decision problem with an associated Bayes rule which has errors α at $\mu = \pm\delta$, and which minimizes our chosen objective function over the set of all decision rules. This Bayes rule can equally be viewed as an optimal stopping rule for our original frequentist problem.

In §3.6 we describe our improved method for objective function $F_2: E(N|\mu=\delta)$. In §3.7 we indicate how to adapt our approach when minimizing other objective functions.

3.6 An Improved Method for the Computation of Optimal One-Sided Group Sequential Tests.

Consider again the clinical trial problem outlined in §3.5 with a maximum of K equally sized groups of n pairs of patients. The random variable X_i $(i = 1, 2, \ldots, Kn)$ denotes the difference in response between the ith pair of patients and is normally distributed with unknown mean μ and known variance σ^2.

Here we wish to make a choice between the decisions :

$$D_1^-: \mu < 0 \qquad \text{and} \qquad D_1^+: \mu > 0.$$

We shall consider a family of Bayes decision theory problems with a common prior distribution for μ given by $\pi(-\delta) = \pi(\delta) = 1/2$ with $\pi(\mu) = 0$ otherwise, and a common cost of sampling function given by $c(-\delta) = c(\delta) = 1$ with $c(\mu) = 0$ otherwise. Individual problems within the family differ in their loss functions, $L(D,\mu)$, which are indexed by a single loss parameter, d (> 0). The general form of $L(D,\mu)$ is given by $L(D_1^-,\delta) = L(D_1^+,-\delta) = d$ with $L(D,\mu) = 0$ otherwise.

Suppose for the moment that d is fixed. Consider a general decision rule for the above problem which we shall denote by $\mathcal{B}$. Because our family of Bayes decision problems are symmetric about $\mu = 0$ we shall only consider decision rules which are symmetric about zero.

So $\mathcal{B}$ is of the general form :

At analysis i $(1 \leq i \leq K-1)$,

if $S_{in} \geq c_i$ stop entering patients on to the trial and make decision D_1^+;

if $S_{in} \leq -c_i$ stop entering patients on to the trial and make decision D_1^-;

if $-c_i < S_{in} < c_i$ enter the next group of n pairs of patients on to the trial.

At analysis K,

if $S_{Kn} > 0$ stop entering patients on to the trial and make decision D_1^+;

if $S_{Kn} < 0$ stop entering patients on to the trial and make decision D_1^- .

The **risk** associated with $\mathcal{B}$, which we shall denote by $r(\mathcal{B},d)$, is defined as the sum of the total expected sampling cost of the trial **plus** the total expected loss through making a wrong decision. That is

$$r(\mathcal{B},d) = c(-\delta)E(N|\mu=-\delta)\pi(-\delta) + c(\delta)E(N|\mu=\delta)\pi(\delta)$$
$$+ d\Pr(D_1^+|\mu=-\delta)\pi(-\delta) + d\Pr(D_1^-|\mu=\delta)\pi(\delta),$$

where $\Pr(D|\mu)$ denotes the probability of making decision D under a treatment difference μ.

As $F_2 = E(N|\mu=\delta) = E(N|\mu=-\delta)$, for symmetric stopping rules, it follows that

$$r(\mathcal{B},d) = \frac{1}{2}\{2F_2 + d\Pr(D_1^+|\mu=-\delta) + d\Pr(D_1^-|\mu=\delta)\}.$$

The Bayes decision rule for our problem, $\mathcal{B}^*(d)$, minimizes this risk over the set of **all** decision rules. Denoting the set of all decision rules by $\mathcal{S}$, we have

$$r(\mathcal{B}^*(d),d) = \min_{\mathcal{B}\in\mathcal{S}}\{r(\mathcal{B},d)\}.$$

We can compute $\mathcal{B}^*(d)$ by dynamic programming (the necessary computations are described in §3.8). Using numerical integration we can calculate the error rates of $\mathcal{B}^*(d)$.

Suppose these errors are given by

$$\Pr(D_1^+|\mu=-\delta) = \Pr(D_1^-|\mu=\delta) = \alpha,$$

then clearly

$$r(\mathcal{B}^*(d),d) = \frac{1}{2}\{2F_2 + 2d\alpha\}$$

$$= F_2 + d\alpha.$$

It follows from the definition of the Bayes rule that there can be no other decision rule with errors α at $\mu = \pm\delta$ which attains a lower value of F_2.

Using a simple bisection search over d we obtain the loss parameter $d^{(\alpha)}$ giving rise to a Bayes decision theory problem with an associated Bayes rule $\mathcal{B}^*(d^{(\alpha)})$ which has errors α at $\mu = \pm\delta$. Now clearly

$$r(\mathcal{B}^*(d^{(\alpha)}),d^{(\alpha)}) = F_2 + d^{(\alpha)}\alpha$$

and, from the definition of the Bayes decision rule, there can be no other decision rule for this problem with a smaller risk, i.e.

$$r(\mathcal{B}^*(d^{(\alpha)}),d^{(\alpha)}) = \min_{\mathcal{B}\in\mathcal{S}}\{r(\mathcal{B},d^{(\alpha)})\}.$$

Moreover there can be no other decision rule with errors α at $\mu = \pm\delta$ which attains a lower value of F_2. Hence $\mathcal{B}^*(d^{(\alpha)})$ minimizes the objective function F_2 over the set of all decision rules with errors α at $\mu = \pm\delta$. By equating decisions D_1^- and D_1^+ with hypotheses H_1^- and H_1^+ respectively we have computed the optimal one-sided group sequential test for our original frequentist problem.

3.7 Applying our Improved Method to Other Objective Functions.

The method described in §3.6 is easily adapted to compute optimal one-sided group sequential tests for other objective functions.

Consider the minimization of F_1: $E(N|\mu=0)$. A suitable family of Bayes decision theory problems has the common prior distribution for μ given by $\pi(-\delta) = \pi(0) = \pi(\delta) = 1/3$ with $\pi(\mu) = 0$ otherwise, and the common cost of sampling function $c(0) = 1$ with $c(\mu) = 0$ otherwise. The general form of the loss function, $L(D,\mu)$, is identical to that given in §3.6, with, in particular, $L(D_1^-,\mu=0) = L(D_1^+,\mu=0) = 0$.

For a fixed loss parameter, d, the **risk** of a given decision rule, $\mathcal{B}$, is denoted by $r(\mathcal{B},d)$ and equals

$$c(0)\,E(N|\mu=0)\,\pi(0) + d\Pr(D_1{}^+|\mu=-\delta)\,\pi(-\delta) + d\Pr(D_1{}^-|\mu=\delta)\,\pi(\delta).$$

That is,

$$r(\mathcal{B},d) = \frac{1}{3}\{F_1 + d\Pr(D_1{}^+|\mu=-\delta) + d\Pr(D_1{}^-|\mu=\delta)\}.$$

The Bayes decision rule for this problem, $\mathcal{B}^*(d)$, minimizes this risk over the set of all decision rules, $\mathcal{S}$, i.e.

$$r(\mathcal{B}^*(d),d) = \min_{\mathcal{B}\in\mathcal{S}}\{r(\mathcal{B},d)\}.$$

Suppose $\mathcal{B}^*(d)$ has errors given by

$$\Pr(D_1{}^+|\mu=-\delta) = \Pr(D_1{}^-|\mu=\delta) = \alpha,$$

then clearly

$$r(\mathcal{B}^*(d),d) = \frac{1}{3}\{F_1 + 2\,d\,\alpha\}$$

and, from the definition of the Bayes rule, there can be no other decision rule with errors α at $\mu=\pm\delta$ which attains a lower value of F_1.

By searching over d (>0) we obtain the loss parameter, $d^{(\alpha)}$, giving a Bayes decision theory problem with an associated Bayes rule, $\mathcal{B}^*(d^{(\alpha)})$, which has errors α at $\mu=\pm\delta$. Clearly

$$r(\mathcal{B}^*(d^{(\alpha)}),d^{(\alpha)}) = \frac{1}{3}\{F_1 + 2\,d^{(\alpha)}\,\alpha\}.$$

Again, from the definition of the Bayes decision rule, there can be no other decision rule for this problem with a smaller risk, i.e.

$$r(\mathcal{B}^*(d^{(\alpha)}),d^{(\alpha)}) = \min_{\mathcal{B}\in\mathcal{S}}\{r(\mathcal{B},d^{(\alpha)})\}$$

Moreover there can be no other decision rule with errors α at $\mu=\pm\delta$ which attains a lower value of F_1. By equating decisions $D_1{}^-$ and $D_1{}^+$ with the acceptance of the hypotheses $H_1{}^-$ and $H_1{}^+$ we obtain the optimal one-sided group sequential test for our original frequentist problem.

For the minimization of F_3: $E(N|\mu=2\delta)$ we consider the family of Bayes decision theory problems with the common prior distribution for μ given by $\pi(-2\delta) = \pi(-\delta) = \pi(\delta) = \pi(2\delta) = 1/4$ with $\pi(\mu) = 0$ otherwise, and the common cost of sampling function given by $c(-2\delta) = c(2\delta) = 1$ with $c(\mu) = 0$ otherwise. The general form of the loss function, $L(D,\mu)$, is identical to that given in §3.6 with, in particular, $L(D_1^-,\mu) = L(D_1^+,\mu) = 0$ for $\mu = \pm 2\delta$.

For a fixed loss parameter, d, the **risk** of a given decision rule, $\mathcal{B}$, is denoted by $r(\mathcal{B},d)$ and equals

$$c(-2\delta)E(N|\mu=-2\delta)\pi(-2\delta) + c(2\delta)E(N|\mu=2\delta)\pi(2\delta)$$
$$+ d\Pr(D_1^+|\mu=-\delta)\pi(-\delta) + d\Pr(D_1^-|\mu=\delta)\pi(\delta).$$

That is

$$r(\mathcal{B},d) = \frac{1}{4}\{2F_3 + d\Pr(D_1^+|\mu=-\delta) + d\Pr(D_1^-|\mu=\delta)\}.$$

The rest of the logic is analogous to that for objective functions F_1 and F_2 given earlier. In particular we search over d (>0) for a loss parameter, $d^{(\alpha)}$, giving a Bayes decision theory problem with associated Bayes rule, $\mathcal{B}^*(d^{(\alpha)})$, which has errors α at $\mu = \pm\delta$. Clearly

$$r(\mathcal{B}^*(d^{(\alpha)}),d^{(\alpha)}) = \frac{1}{4}\{2F_3 + 2d^{\alpha}\alpha\} = \frac{1}{2}\{F_3 + d^{(\alpha)}\alpha\}.$$

From the definition of the Bayes rule, there can be no other decision rule for this problem with a smaller risk. Moreover there can be no other decision rule with errors α at $\mu = \pm\delta$ which attains a lower value of F_3. By equating decisions D_1^- and D_1^+ with the acceptance of the hypotheses H_1^- and H_1^+ we obtain the optimal one-sided group sequential test for our original frequentist problem.

Jennison (1987) considered the minimization of F_4 because he wanted to compute optimal tests which were optimal, or near optimal, over a range of parameter values. Instead of considering the minimization of F_4 we will consider the minimization of $E(N|\mu)$ integrated over a normal density with mean 0 and variance δ^2. Denoting this objective function by F_5, we have:

$$F_5: \int E(N|\mu)\delta^{-1}\varphi(\mu/\delta)\,d\mu.$$

A suitable family of Bayes decision theory problems for the minimization of F_5 has the common prior distribution:

$$\pi(\mu) = \begin{cases} 1/3 & \text{if } \mu = \pm\delta \\ (1/3)\,\delta^{-1}\varphi(\mu/\delta) & \text{otherwise} \end{cases}$$

and the common cost of sampling function :

$$c(\mu) = \begin{cases} 1 \text{ if } \mu \neq \pm\delta \\ 0 \text{ otherwise.} \end{cases}$$

Again the general form of the loss function, $L(D,\mu)$, is identical to that given in §3.6, with, in particular, $L(D,\mu) = 0$ for $\mu \neq \pm\delta$.

For a fixed loss parameter, d, the **risk** of a given decision rule, $\mathcal{B}$, is denoted by $r(\mathcal{B},d)$ and is equal to

$$r(\mathcal{B},d) = \int c(\mu)\,E(N|\mu)\,\pi(\mu)\,d\mu +$$

$$d\,\Pr(D_1^+|\mu=-\delta)\,\pi(-\delta) + d\,\Pr(D_1^-|\mu=\delta)\,\pi(\delta).$$

It follows that

$$r(\mathcal{B}(d)) = \frac{1}{3}\left\{\int_{\mu} E(N|\mu)\,\delta^{-1}\,\varphi(\mu/\delta)\,d\mu + d\,\Pr(D_1^+|\mu=-\delta) + d\,\Pr(D_1^-|\mu=\delta)\right\}$$

$$= \frac{1}{3}\{F_5 + d\,\Pr(D_1^+|\mu=-\delta) + d\,\Pr(D_1^-|\mu=\delta)\}.$$

The rest of the logic is analogous to that for objective functions F_1 and F_2. In particular we search over d for a loss parameter, $d^{(\alpha)}$, giving a Bayes decision theory problem with associated Bayes rule, $\mathcal{B}^*(d^{(\alpha)})$, which has errors α at $\mu = \pm\delta$. Clearly

$$r(\mathcal{B}^*(d^{(\alpha)}),d^{(\alpha)}) = \frac{1}{3}\{F_5 + 2d^{(\alpha)}\alpha\}.$$

From the definition of the Bayes rule, there can be no other decision rule for this problem with a smaller risk. Moreover there can be no other decision rule with errors α at $\mu = \pm\delta$ which attains a lower value of F_5. By equating decisions D_1^- and D_1^+ with the acceptance of the hypotheses H_1^- and H_1^+ we obtain the optimal one-sided group sequential test for our original frequentist problem.

Clearly our improved method for computing optimal one-sided group sequential tests is easily extended to other objective functions not considered here. We now go on to consider in detail the dynamic programming algorithm for our method.

3.8 The Dynamic Programming Algorithm.

For any given Bayes decision theory problem we compute the Bayes decision rule by dynamic programming. In this section we describe the dynamic programming algorithm for a general problem. Throughout we shall assume that the Bayes rule is **monotone**. That is at analysis i $(i = 1, 2, \ldots, K-1)$ it is optimal to make decision ${D_1}^+$ if $S_{in} \geq c_i$ and decision ${D_1}^-$ if $S_{in} \leq -c_i$, while for $-c_i < S_{in} < c_i$ it is optimal to sample the next group of n observations. Further, at analysis K, it is optimal to make decision ${D_1}^+$ if $S_{Kn} \geq 0$ and decision ${D_1}^-$ if $S_{Kn} \leq 0$.

Lai (1973) proved that optimal tests for objective function F_1 are monotone, while Brown, Cohen & Strawdermann (1981) proved the monotonicity of optimal tests for F_2. Although we have no formal proofs to demonstrate the monotonicity of optimal tests for F_3 and F_5, numerical checks support our assumption that these objective functions are monotone.

Suppose we have a maximum of K groups of n observations for choosing between the decisions

$$ {D_1}^-: \mu < 0 \quad \text{and} \quad {D_1}^+: \mu > 0. $$

Further suppose that our family of problems has a common prior distribution denoted by $\pi(\mu)$ which is defined over some parameter space M, and a common cost of sampling function denoted by $c(\mu)$. Finally we shall suppose that our loss function is denoted by $L(D,\mu)$ and that it is of the same general form as that given in §3.6.

Suppose the loss parameter d is fixed. Letting $p^{(i)}(\mu|x)$ denote the current posterior distribution of μ at analysis i $(1 \le i \le K)$, given that $S_{in} = x$, the loss from making decision D_1^+ equals

$$dp^{(i)}(-\delta|x)$$

and the loss from making decision D_1^- equals

$$dp^{(i)}(\delta|x).$$

If we let $\gamma^{(i)}(x)$ denote the minimum loss from stopping at stage i $(i = 1, 2, \ldots, K)$ and making a definite decision, it follows that

$$\gamma^{(i)}(x) = \min \{dp^{(i)}(-\delta|x), dp^{(i)}(\delta|x)\}.$$

Clearly

$$\gamma^{(i)}(x) = \begin{cases} dp^{(i)}(-\delta|x) & \text{for } x > 0 \\ dp^{(i)}(\delta|x) & \text{for } x < 0. \end{cases}$$

Further, let $\beta^{(i)}(x)$ be the minimum additional risk from sampling the $(i+1)$st group $(i = 1, 2, \ldots, K-1)$ and then proceeding optimally. Denoting the c.d.f. of $S_{(i+1)n}$ given that $S_{in} = x$ by $F^{(i+1)}(S_{(i+1)n}|x)$, we have

$$\beta^{(K-1)}(x) = n \sum_{\mu \in M} c(\mu) p^{(K-1)}(\mu|x) + \int_{S_{Kn}} \gamma^{(K)}(S_{Kn})\, dF^{(K)}(S_{Kn}|x)$$

and, for $i < K-1$,

$$\beta^{(i)}(x) = n \sum_{\mu \in M} c(\mu) p^{(i)}(\mu|x) +$$

$$\int_{S_{(i+1)n}} \min\{\beta^{(i+1)}(S_{(i+1)n}), \gamma^{(i+1)}(S_{(i+1)n})\}\, dF^{(i+1)}(S_{(i+1)n}|x)$$

where the summation signs in the above equations are replaced by mixture of summations and integrals for objective functions such as F_5.

We compute the critical values of $\mathcal{B}^*(d)$ by starting at the Kth analysis and working back.

At analysis K it is optimal to make decision D_1^+ for all $S_{Kn} = x$ such that

$$dp^{(K)}(-\delta|x) < dp^{(K)}(\delta|x) \tag{3.8.1}$$

and to make decision D_1^- for all $S_{Kn} = x$ such that

$$dp^{(K)}(\delta|x) < dp^{(K)}(-\delta|x). \tag{3.8.2}$$

Clearly this is equivalent to making decision D_1^+ for $x>0$ and decision D_1^- for $x<0$.

At analysis $K-1$ it is optimal to make decision D_1^+ for all $S_{(K-1)n}=x$ (≥ 0) such that

$$dp^{(K-1)}(-\delta|x) < \beta^{(K-1)}(x) \tag{3.8.3}$$

and to decide to sample the Kth group of observations and then to proceed optimally for all x (≥ 0) such that

$$\beta^{(K-1)}(x) < dp^{(K-1)}(-\delta|x). \tag{3.8.4}$$

Here $\beta^{(K-1)}(x)$ is equal to

$$\beta^{(K-1)}(x) = n \sum_{\mu \in M} c(\mu) p^{(K-1)}(\mu|x)$$

$$+ d\{\Pr_{-\delta}(S_{Kn} \geq 0 \mid x)\, p^{(K-1)}(-\delta|x)\} + d\{\Pr_{\delta}(S_{Kn} \leq 0 \mid x)\, p^{(K-1)}(\delta|x)\}.$$

We use the bisection method to obtain the $(K-1)$st critical value, c_{K-1}, as the solution, for $x \geq 0$, of equation (3.8.5)

$$dp^{(K-1)}(-\delta|x) = \beta^{(K-1)}(x). \tag{3.8.5}$$

Clearly it is possible that equation (3.8.5) does not possess a solution for $x \geq 0$. To avoid our algorithm getting in to computational difficulties we can build a simple check in to our program. The check tests whether inequality (3.8.4) holds for $x=0$. If this is not the case we set $c_{K-1}=0$ and go back to the $(K-2)$nd analysis. Similar checks can be built in at other analyses.

The symmetry inherent in our problem makes it optimal to choose D_1^- if $S_{(K-1)n} \leq -c_{K-1}$ and to continue sampling if $-c_{K-1} < S_{(K-1)n} \leq 0$.

At analysis $K-2$ it is optimal to make decision D_1^+ for all $S_{n(K-2)}=x$ (≥ 0) such that

$$dp^{(K-2)}(-\delta|x) < \beta^{(K-2)}(x)$$

and to decide to enter the $(K-1)$st group of patients on to the trial and then to proceed optimally for all x (≥ 0) such that

$$\beta^{(K-2)}(x) < dp^{(K-2)}(-\delta|x).$$

Here $\beta^{(K-2)}(x)$ is given by

$$\beta^{(K-2)}(x) = n \sum_{\mu \in M} c(\mu) p^{(K-2)}(\mu|x) +$$

$$d\{\Pr_{-\delta}(S_{(K-1)n} \geq c_{K-1}|x)\, p^{(K-2)}(-\delta|x)\} + d\{\Pr_{\delta}(S_{(K-1)n} \leq -c_{K-1}|x)\, p^{(K-2)}(\delta|x)\}$$

$$+ \int_{-c_{K-1}}^{c_{K-1}} \beta^{(K-1)}(S_{(K-1)n})\, dF^{(K-1)}(S_{(K-1)n}|x),$$

Again the bisection method gives the $(K-2)$nd critical value, c_{K-2}, as the solution, for $x \geq 0$, of equation (3.8.6)

$$d\, p^{(K-2)}(-\delta|x) = \beta^{(K-2)}(x). \tag{3.8.6}$$

The symmetrical nature of our problem makes it optimal to choose D_1^- if $S_{n(K-2)} \leq -c_{K-2}$ and to continue sampling if $-c_{K-2} < S_{n(K-2)} \leq 0$.

We work back to the first analysis in a similar fashion. The error probabilities of the resulting decision rule are then computed and a test for convergence conducted. If convergence has not been achieved a new value of d is obtained from the bisection method. The above algorithm is computationally very fast and efficient.

3.9 Results and Discussion.

Tables 3.1-3.8 give the minima of objective functions F_1, F_2, F_3 and F_5 for group sequential tests with equally sized groups, $\alpha = 0.01$ and 0.05, K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200 and t (the ratio of the maximum sample size of the sequential test to the corresponding fixed sample size, N_f) equal to 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6. The minima are expressed as percentages of N_f and are independent of the choice of δ and σ^2 as is verified in Appendix 3.2. In each table the minimum over t for fixed K is shown in bold type.

Tables 3.1 and 3.2 refer to the minimization of objective function F_1: $E(N|\mu{=}0)$ for $\alpha = 0.01$ and 0.05 respectively. The minimization of F_1 is known in the literature as the Kiefer-Weiss problem. For any stopping rule symmetric about zero $E(N|\mu)$ is maximized when $\mu = 0$. Hence the Kiefer-Weiss

problem is essentially a minimax problem.

Lorden (1976), building on the work of Lai (1973), used techniques very similar to our own to determine the minimum of F_1 up to a small discretization error. For $\alpha = 0.01$ this minimum is equal to 71.1% of the fixed sample size and occurs with continuous monitoring of the data and $t=2.33$. For $\alpha=0.05$ the minimum of F_1 is equal to 73.3% of the fixed sample size. It also occurs with continuous monitoring of the data but this time with $t = 3.37$.

Clearly, in the context of a clinical trial, continuous monitoring of the data is likely to prove impractical. As we have already seen, Jennison (1987) considered minimizing F_1 over feasible group sequential tests. Tables 3.1 and 3.2 extend the results of Jennison by considering designs with $K = 4$, $K > 10$ and $t = 1.01$.

In discussing Table 3.1 we begin by considering group sequential tests with $K=2$. Compared with the fixed sample size test satisfying the same error constraints, each of the 2 group tests lead to savings in F_1. The largest of these savings occurs with $t = 1.1$ and is equal to 10.6% of N_f. Note that even if logistical considerations restrict us to designs with $t = 1.01$ the expected saving is 6% of the fixed sample size. In a clinical trial with perhaps hundreds of patients this could represent an important saving in both financial and human resources.

As might be anticipated the expected gains in efficiency over the fixed sample size test are even more impressive when $K=3$. The largest saving occurs with $t=1.15$ and is equal to 14.8% of the fixed sample size. When $t=1.01$ the expected saving equals 9% of N_f.

Pocock (1982) suggested that in practice it would be unlikely for a clinical trial to be designed with more than 5 analyses. The most efficient 5 group design in our table occurs with $t = 1.2$ and on average leads to an 18.8% saving over the fixed sample size test. This compares favourably with the most efficient 200 group test in the table and with the approximately optimal test of Lorden (1976).

The most efficient test in Table 3.1 has a maximum of 200 groups and $t = 1.5$. For this test $E(N|\mu=0)$ is only 0.2% of the fixed sample size greater than for Lorden's (1976) test.

From the table it is clear that, for fixed t, the rate of gain in efficiency in F_1 depreciates as K increases. For example, when $t = 1.5$ doubling the maximum number of analyses from 100 to 200 produces a mere 0.2% of N_f gain in

efficiency. It is also clear from the table that for each value of K the minima of F_1 are a U-shaped function of t. This observation was initially made by Jennison (1987).

Table 3.1. Minima of $F_1 : E(N|\mu=0)$ expressed as percentages of the fixed sample size, N_f, for $\alpha = 0.01$, t = 1.01, 1.05 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	94.0	90.3	**89.4**	89.5	90.0	91.5	93.3	95.0	96.6
3	91.0	86.8	85.5	**85.2**	85.3	86.1	87.1	88.0	88.8
4	89.4	84.8	83.3	82.8	**82.8**	83.2	84.0	84.7	85.2
5	88.3	83.6	81.9	81.3	**81.2**	81.5	82.0	82.6	83.1
10	86.1	81.0	79.1	78.3	**77.9**	77.9	78.1	78.3	78.6
15	85.3	80.1	78.1	77.3	76.9	**76.7**	76.8	76.9	77.1
20	84.8	79.7	77.7	76.8	76.3	**76.1**	76.1	76.3	76.4
30	84.4	79.2	77.2	76.3	75.8	**75.5**	75.5	75.6	75.7
50	84.1	78.9	76.8	75.9	75.4	75.1	**75.0**	75.1	75.1
100	83.9	78.7	76.5	75.6	75.1	74.7	74.7	**74.7**	74.7
200	83.8	78.5	76.4	75.4	74.9	74.6	74.5	**74.5**	74.5

Many of the comments made concerning Table 3.1 are equally applicable to Table 3.2. Again the largest expected gains in efficiency come from adopting a 2 group sequential test rather than a single sample test. For example with $K=2$ and $t=1.15$ an expected gain in efficiency of 13% of the fixed sample size is attained.

By considering a test with 10 groups and $t=1.4$ an expected gain in efficiency of 25.1% of the fixed sample size is obtained. This is only 3.8% of N_f less than that for the Lorden (1976) test with $\alpha=0.05$. The optimum test in Table 3.2 has $K=200$ and $t=1.6$ which is a mere 0.2% of the fixed sample size less than for Lorden's test.

Table 3.2. Minima of $F_1: E(N|\mu=0)$ expressed as percentages of the fixed sample size, N_f, for $\alpha=0.05$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	93.4	88.9	87.3	**87.0**	87.2	88.3	90.0	91.7	93.6
3	90.6	85.6	83.5	82.7	**82.4**	82.6	83.3	84.2	85.1
4	88.9	83.7	81.5	80.5	80.1	**80.0**	80.4	80.9	81.6
5	87.9	82.5	80.2	79.1	78.6	**78.4**	78.6	79.0	79.5
10	85.7	80.0	77.5	76.2	75.6	75.0	**74.9**	75.0	75.3
15	84.9	79.1	76.5	75.3	74.5	73.9	**73.7**	73.7	73.8
20	84.5	78.7	76.1	74.8	74.0	73.3	73.1	**73.1**	73.1
30	84.1	78.2	75.6	74.3	73.5	72.7	72.5	**72.4**	72.4
50	83.7	77.9	75.2	73.9	73.1	72.3	72.0	71.9	**71.9**
100	83.5	77.6	75.0	73.6	72.8	72.0	71.6	71.5	**71.5**
200	83.4	77.4	74.8	73.4	72.6	71.8	71.4	71.3	**71.3**

Tables 3.3 and 3.4 refer to the minimization of objective function $F_2: E(N|\mu=\delta)$ for $\alpha=0.01$ and 0.05 respectively. As was seen in §3.4 the minimization of F_2 has received a great deal of attention in the literature. The SPRT of Wald (1947) minimizes F_2 over the set of feasible tests, but requires continuous monitoring of the data and no upper bound on the maximum sample size. With $\alpha=0.01$, the expected sample size under $\mu=\delta$ for the SPRT is just 41.6% of the fixed sample size, while with $\alpha=0.05$ it is 49% of the fixed. Jennison (1987) considered the minimization of F_2 over feasible group sequential tests with finite maximum sample sizes. Tables 3.3 and 3.4 extend the results of Jennison by including the minima of F_2 for tests with $K=4$, $K>10$ and $t=1.01$.

Table 3.3. Minima of $F_2 : E(N|\mu=\delta)$ expressed as percentages of the fixed sample size, N_f, for $\alpha = 0.01$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	72.3	66.7	**65.8**	66.3	67.4	70.4	73.9	77.8	82.0
3	67.1	60.4	58.1	57.3	**57.0**	57.6	58.8	60.3	62.2
4	64.3	57.7	55.2	54.0	53.5	**53.2**	53.6	54.2	55.1
5	62.5	56.0	53.5	52.2	51.6	**51.1**	51.1	51.4	51.9
10	59.1	52.6	49.9	48.6	47.8	47.0	46.7	**46.7**	46.8
15	58.1	51.5	48.8	47.4	46.6	45.7	45.3	**45.1**	45.1
20	57.5	50.9	48.2	46.8	46.0	45.0	44.6	44.4	**44.3**
30	57.0	50.4	47.7	46.3	45.4	44.4	43.9	43.6	**43.5**
50	56.6	50.0	47.3	45.8	44.9	43.9	43.3	43.0	**42.9**
100	56.4	49.7	47.0	45.5	44.6	43.5	43.0	42.6	**42.4**
200	56.2	49.6	46.8	45.4	44.4	43.4	42.8	42.4	**42.2**

Consider Table 3.3. Again the largest expected gains in efficiency are obtained by going from a single sample test to a group sequential test with a maximum of 2 groups. These gains are very impressive with the largest occurring when $t = 1.1$ and being equal to 34.2% of the fixed sample size. Even if logistical considerations limit us to a test with $K = 2$ and $t = 1.01$ the expected gain in efficiency is more than one-quarter of the fixed sample size.

With $K = 200$ and $t = 1.6$ we are within 0.6% of the fixed sample size from the SPRT minimum. Indeed with as few as 10 groups and with $t = 1.5$ our optimal test is just 5.1% of N_f from the SPRT minimum. Clearly the differences in efficiency between tests with a maximum of 10 groups and the SPRT are small, and so tests likely to be used in practice are impressively efficient.

Table 3.4. Minima of $F_2: E(N|\mu=\delta)$ expressed as percentages of the fixed sample size, N_f, for $\alpha = 0.05$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	80.9	74.5	72.8	**72.7**	73.2	75.3	78.1	81.2	84.7
3	76.3	69.3	66.5	65.3	**64.8**	64.8	65.6	66.7	68.2
4	73.8	66.8	63.8	62.4	61.6	**61.0**	61.0	61.4	62.1
5	72.2	65.2	62.2	60.7	59.8	59.0	**58.7**	58.8	59.2
10	69.1	62.1	59.0	57.4	56.3	55.2	54.6	54.4	**54.3**
15	68.1	61.0	57.9	56.2	55.2	53.9	53.3	53.0	**52.8**
20	67.6	60.5	57.4	55.7	54.6	53.3	52.6	52.2	**52.0**
30	67.2	60.1	56.9	55.2	54.0	52.7	52.0	51.5	**51.3**
50	66.8	59.7	56.5	54.7	53.6	52.2	51.5	51.0	**50.7**
100	66.5	59.4	56.2	54.4	53.3	51.9	51.1	50.6	**50.2**
200	66.3	59.2	56.0	54.3	53.1	51.7	50.9	50.4	**50.0**

The most efficient group sequential test in Table 3.4 has a maximum of 200 groups and $t = 1.6$. For this test $E(N|\mu=\delta)$ is within 1% of the fixed sample size from that for the SPRT. Again the loss in efficiency from considering tests with 10 groups or fewer is small. For example with $K = 10$ and $t = 1.6$ the optimal test is within 5% of the fixed sample size from the SPRT minimum.

An important difference between the optimal tests for F_2 given in Tables 3.3 and 3.4 and the optimal tests for F_1 given in Tables 3.1 and 3.2 concerns the value of t at which the objective function is minimized for fixed K. For $\alpha = 0.05$ and $K \geq 10$, F_2 is minimized when $t = 1.6$. A similar pattern emerges for F_2 when $\alpha = 0.01$, with, for $K \geq 20$, minima occurring when $t = 1.6$. The corresponding minima for F_1 tend to occur at lower values of t.

To demonstrate just how efficient our method is, we note that to obtain the optimal test with $t = 1.6$ and $K = 2$ required just 2 seconds of CPU time on a Sun-4 computer. For $K = 3$ and $t = 1.6$, 4.9 seconds of CPU was required, while

for $K = 10$ and $t = 1.6$, 167.2 seconds of CPU time was needed. Compare this with the original approach of Jennison (1987) which required something in the order of 10 hours of CPU time to compute this last test.

The minimization of F_3: $E(N|\mu=2\delta)$ was first considered by Jennison (1987). Optimal tests for this objective function are designed to ensure that a trial is stopped early when treatment differences are large. These tests are not very robust, however, and, on average, may require more patients than the corresponding fixed sample size test when $|\mu|$ is small.

Table 3.5. Minima of $F_3: E(N|\mu=2\delta)$ expressed as percentages of the fixed sample size, N_f, for $\alpha = 0.01$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

					t				
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	**52.2**	53.1	55.3	57.7	60.1	65.0	70.0	75.0	80.0
3	41.4	**38.8**	39.0	40.0	41.2	44.1	47.1	50.3	53.5
4	38.2	33.9	**33.0**	33.0	33.4	34.8	36.7	38.7	40.9
5	36.6	31.8	30.3	**29.8**	29.8	30.3	31.4	32.6	34.1
10	33.2	28.4	26.5	25.5	25.0	24.4	**24.3**	24.3	24.5
15	32.0	27.2	25.3	24.3	23.7	23.1	22.8	22.6	**22.6**
20	31.5	26.6	24.7	23.7	23.1	22.4	22.0	21.9	**21.8**
30	30.9	26.1	24.1	23.1	22.5	21.7	21.3	21.1	**21.0**
50	30.5	25.7	23.7	22.6	22.0	21.2	20.7	20.5	**20.3**
100	30.3	25.4	23.4	22.3	21.6	20.8	20.3	20.0	**19.9**
200	30.1	25.2	23.2	22.2	21.5	20.6	20.1	19.8	**19.6**

There are no analytical results in the literature referring to the minimization of F_3. Tables 3.5 and 3.6, which give the minima of F_3 for $\alpha = 0.01$ and 0.05 and the same designs as in Tables 3.1-3.4, enable us to obtain good upper bounds on the overall minima of F_3. For instance, from Table 3.5, the best test has a maximum of 200 groups and $t = 1.6$, and, on average, requires only 19.6% of the

corresponding fixed sample size.

Tables 3.5 and 3.6 provide the most impressive justification for the adoption of group sequential tests. With $K=2$, $t=1.01$ and $\alpha=0.01$ our optimal test leads to an average gain in efficiency of 47.8% N_f. For the same problem but with $\alpha=0.05$ the average gain in efficiency is 40.3% of the fixed sample size. As with objective functions F_1 and F_2 gains in efficiency increase with K. However for K greater than 10 the rate of gain in efficiency is small.

Table 3.6. Minima of $F_3: E(N|\mu=2\delta)$ expressed as percentages of the fixed sample size, N_f, for $\alpha=0.05$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	59.7	**57.0**	57.9	59.5	61.5	65.9	70.5	75.3	80.2
3	53.0	46.9	45.4	**45.3**	45.7	47.4	49.7	52.3	55.0
4	50.2	43.5	41.1	40.2	**39.9**	40.2	41.2	42.5	44.1
5	48.6	41.7	39.1	37.8	37.2	**36.8**	37.1	37.7	38.6
10	45.1	38.4	35.6	34.0	33.1	31.9	31.3	31.0	**30.8**
15	44.0	37.3	34.4	32.9	31.8	30.6	29.8	29.4	**29.1**
20	43.4	36.7	33.8	32.3	31.2	29.9	29.2	28.7	**28.3**
30	42.9	36.2	33.3	31.7	30.6	29.3	28.5	27.9	**27.5**
50	42.5	35.8	32.9	31.3	30.2	28.8	27.9	27.3	**26.9**
100	42.2	35.4	32.6	30.9	29.8	28.4	27.5	26.9	**26.5**
200	42.0	35.2	32.4	30.8	29.7	28.2	27.3	26.7	**26.3**

Tables 3.7 and 3.8 give the minima of objective function $F_5: \int E(N|\mu)\delta^{-1}\varphi(\mu/\delta)d\mu$ for $\alpha=0.01$ and 0.05 respectively. The motivation for considering optimal tests for F_5 lies in the desire to compute tests which are optimal or close to optimal over the entire parameter space for μ. As we shall see in §3.11 these optimal tests are indeed very robust.

Table 3.7. Minima of $F_5: \int E(N|\mu)\delta^{-1}\varphi(\mu/\delta)\,d\mu$ expressed as percentages of the fixed sample size, N_f, for $\alpha = 0.01$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	78.6	74.4	**73.9**	74.4	75.4	77.9	80.9	84.0	87.2
3	74.0	68.7	67.0	**66.6**	66.7	67.5	68.9	70.4	72.0
4	71.6	66.1	64.2	63.5	**63.2**	63.5	64.2	65.1	66.1
5	70.2	64.6	62.6	61.7	61.4	**61.4**	61.9	62.5	63.1
10	67.2	61.5	59.3	58.4	57.9	**57.5**	57.6	57.8	58.1
15	66.2	60.5	58.3	57.2	56.7	56.2	**56.2**	56.3	56.4
20	65.8	60.0	57.7	56.7	56.1	55.6	**55.5**	55.6	55.7
30	65.3	59.5	57.2	56.1	55.5	55.0	**54.8**	54.8	54.9
50	64.9	59.1	56.8	55.7	55.1	54.5	54.3	**54.3**	54.3
100	64.6	58.8	56.5	55.4	54.7	54.1	53.9	**53.9**	53.9
200	64.5	58.7	56.4	55.2	54.6	54.0	53.7	53.7	**53.6**

Many of the comments made concerning objective functions F_1, F_2 and F_3 apply equally here. Again the largest gains in efficiency occur when going from a fixed sample size test to a 2 group sequential test. Also the average gain in efficiency is still substantial if we limit attention to designs with only 2 groups and a maximum sample size only 1% greater than the fixed sample size. With $\alpha = 0.01$ this gain is 21.4% of the fixed sample size, while with $\alpha = 0.05$ the gain is 16.7% of N_f. Most of the gains in efficiency are obtained with tests with 10 groups or less. As with objective function F_3 there are no analytical results or approximations in the literature to the overall minima of F_5. Tables 3.7 and 3.8 enable us to obtain fairly accurate approximations to these minima.

For instance with $\alpha = 0.01$ it would appear that even with continuous monitoring of the data the overall minimum would be no less than 53% of the fixed sample size. With $\alpha = 0.05$ the overall minimum should be approximately 56% of N_f.

Table 3.8. Minima of $F_5: \int E(N|\mu)\delta^{-1}\varphi(\mu/\delta)$ expressed as percentages of the fixed sample size, N_f, for $\alpha = 0.05$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	83.3	78.1	**76.8**	76.8	77.4	79.4	81.9	84.7	87.7
3	79.4	73.3	71.0	70.0	**69.7**	70.0	70.9	72.2	73.6
5	75.9	69.6	67.0	65.7	65.0	**64.5**	64.6	64.9	65.4
10	73.1	66.7	63.9	62.5	61.7	60.9	**60.6**	60.6	60.7
15	72.7	65.7	62.9	61.4	60.6	59.7	59.3	**59.2**	59.2
20	71.7	65.2	62.4	60.9	60.0	59.0	58.6	**58.5**	58.5
30	71.3	64.8	61.9	60.4	59.5	58.5	58.0	57.8	**57.7**
50	70.9	64.4	61.5	60.0	59.0	58.0	57.5	57.2	**57.1**
100	70.6	64.1	61.2	59.7	58.7	57.6	57.1	56.8	**56.7**
200	70.5	64.0	61.1	59.5	58.6	57.5	56.9	56.6	**56.5**

3.10 The Loss Functions for the Optimal Tests.

Tables 3.9 and 3.10 give values of $d/\{cN_f\}$, where c is the cost of sampling an observation, for the optimal designs for objective function F_2 with $\alpha = 0.01$ and 0.05 respectively. The entries in each table are given to one decimal place and are independent of δ and σ^2.

Table 3.9. Values of $d/\{c N_f\}$ corresponding to the minima of objective function F_2 with $\alpha = 0.01$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	151.2	33.4	17.5	12.1	9.3	6.5	5.0	4.1	3.4
3	153.4	41.1	24.0	17.9	14.7	11.3	9.4	8.2	7.2
4	146.7	41.0	25.0	19.2	16.2	13.0	11.3	10.2	9.4
5	144.2	40.3	24.8	19.3	16.4	13.5	12.0	11.1	10.4
10	142.9	39.5	24.2	18.8	16.1	13.4	12.1	11.4	10.9
15	142.4	39.4	24.1	18.8	16.0	13.3	12.0	11.3	10.8
20	142.1	39.4	24.1	18.8	16.0	13.3	12.0	11.3	10.8
30	142.0	39.3	24.1	18.7	16.0	13.3	12.0	11.3	10.8
50	141.8	39.3	24.1	18.7	16.0	13.3	12.0	11.3	10.8
100	141.8	39.3	24.1	18.7	16.0	13.3	12.0	11.3	10.8
200	141.8	39.3	24.1	18.8	16.0	13.3	12.0	11.3	10.8

The pattern in both tables is similar with little variation in $d/\{cN_f\}$ over K for fixed t, but entries increasing rapidly as t decreases towards 1.

Medical researchers are, understandably, reluctant to specify decision theory loss functions for clinical trials; however by considering tables such as Table 3.9 and Table 3.10 ethically appropriate designs can be arrived at. For example consider a hypothesis testing problem with $\alpha = 0.05$ and fixed sample size, $N_f = 100$. If the researcher feels that the loss through making a wrong decision is 1 000 times more expensive than the cost of admitting a further pair of patients on to the trial then he should choose a design with $d/\{cN_f\}$ equal to 10. From Table 3.10 it can be seen that this corresponds to a group sequential experiment with t approximately equal to 1.1.

Alternatively the researcher might have arrived at values of K, t and α and wish to simply check that the associated value of $d/\{cN_f\}$ is reasonable. He might find, especially with $\alpha = 0.05$ and $t \geq 1.3$, that the losses he is implicitly

working with are unreasonable. To overcome this problem the experimenter should decrease α.

Table 3.10. Values of $d/\{c\,N_f\}$ corresponding to the minima of objective function F_2 with $\alpha = 0.05$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	57.5	15.7	9.2	6.8	5.6	4.3	3.6	3.1	2.8
3	56.7	17.6	11.0	8.6	7.2	5.8	5.0	4.5	4.1
4	55.5	17.5	11.3	8.9	7.6	6.3	5.6	5.1	4.8
5	55.2	17.3	11.2	8.9	7.7	6.4	5.8	5.3	5.0
10	55.0	17.1	11.0	8.8	7.6	6.4	5.8	5.4	5.2
15	54.9	17.1	11.0	8.8	7.6	6.4	5.8	5.4	5.2
20	54.8	17.1	11.0	8.8	7.6	6.4	5.8	5.4	5.2
30	54.7	17.1	11.0	8.8	7.6	6.4	5.8	5.4	5.2
50	54.7	17.1	11.0	8.8	7.6	6.4	5.8	5.4	5.2
100	54.7	17.1	11.0	8.8	7.6	6.4	5.8	5.4	5.2
200	54.7	17.1	11.0	8.8	7.6	6.4	5.8	5.4	5.2

As can be seen from Table 3.9, values of $d/\{cN_f\}$ for $\alpha = 0.01$ are universally higher than in Table 3.10.

3.11 Examples.

In this section we give four examples of optimal one-sided group sequential tests. All of the examples are based on the same hypothesis testing problem which we now describe.

Suppose $X_1, X_2, \ldots, X_{Kn}$ are independent normal random variables with unknown mean μ and unit variance. A maximum of 5 groups of 10 observations are available for testing

$$H_1^-: \mu = -0.25 \quad vs \quad H_1^+: \mu = 0.25$$

with error rates

$$\Pr(\mathcal{A}_1^+ \mid \mu = -0.25) = \Pr(\mathcal{A}_1^- \mid \mu = 0.25) = 0.05.$$

The fixed sample size test for this problem requires 44 observations.

We shall consider the four optimal group sequential tests which minimize the objective functions F_1, F_2, F_3 and F_5. To ease notation we shall denote these four tests by T_1, T_2, T_3 and T_5 respectively. Table 3.11 gives the critical values and the attained values of F_1, F_2, F_3 and F_5 for the four optimal tests.

Table 3.11. The critical values and the attained values of F_1, F_2, F_3 and F_5 for the four optimal group sequential tests T_1, T_2, T_3 and T_5.

Test	Critical Values					Objective Function			
	c_1	c_2	c_3	c_4	c_5	F_1	F_2	F_3	F_5
T_1	6.243	5.141	4.010	2.727	0.0	34.2	26.7	18.1	28.6
T_2	5.274	5.050	4.623	3.697	0.0	34.6	26.2	16.9	28.4
T_3	4.586	5.496	6.021	5.663	0.0	36.5	27.1	16.3	29.5
T_5	5.431	5.121	4.441	3.276	0.0	34.4	26.3	17.1	28.4

As can be seen from the table, the stopping rules for T_2 and T_5 are very similar. As would be expected the stopping rule for T_1 is conservative at the first analysis. Conversely the stopping rule for T_3 is very liberal at the first analysis.

The similarity of the stopping rules for T_2 and T_5 is highlighted in the attained values of F_1, F_2, F_3 and F_5. Of course T_2 is the best of the four tests in terms of minimizing F_2, although T_5 requires just 0.1 observations more on average. Indeed all of the optimal tests are, on average, within 1 observation of

the overall minimum. Compare this with the fixed sample size test which require 17.8 observations more than T_2 on average.

In terms of minimizing F_5 test T_5 is optimal, although to 1 decimal place there is no difference between T_5 and T_2. Test T_1 is close to optimal for this objective function, while T_3 requires 1.1 more observations than T_2 and T_5 on average. All 4 group sequential tests can be seen to be substantial improvements on the fixed sample size test which requires 15.6 more observations than T_5 on average.

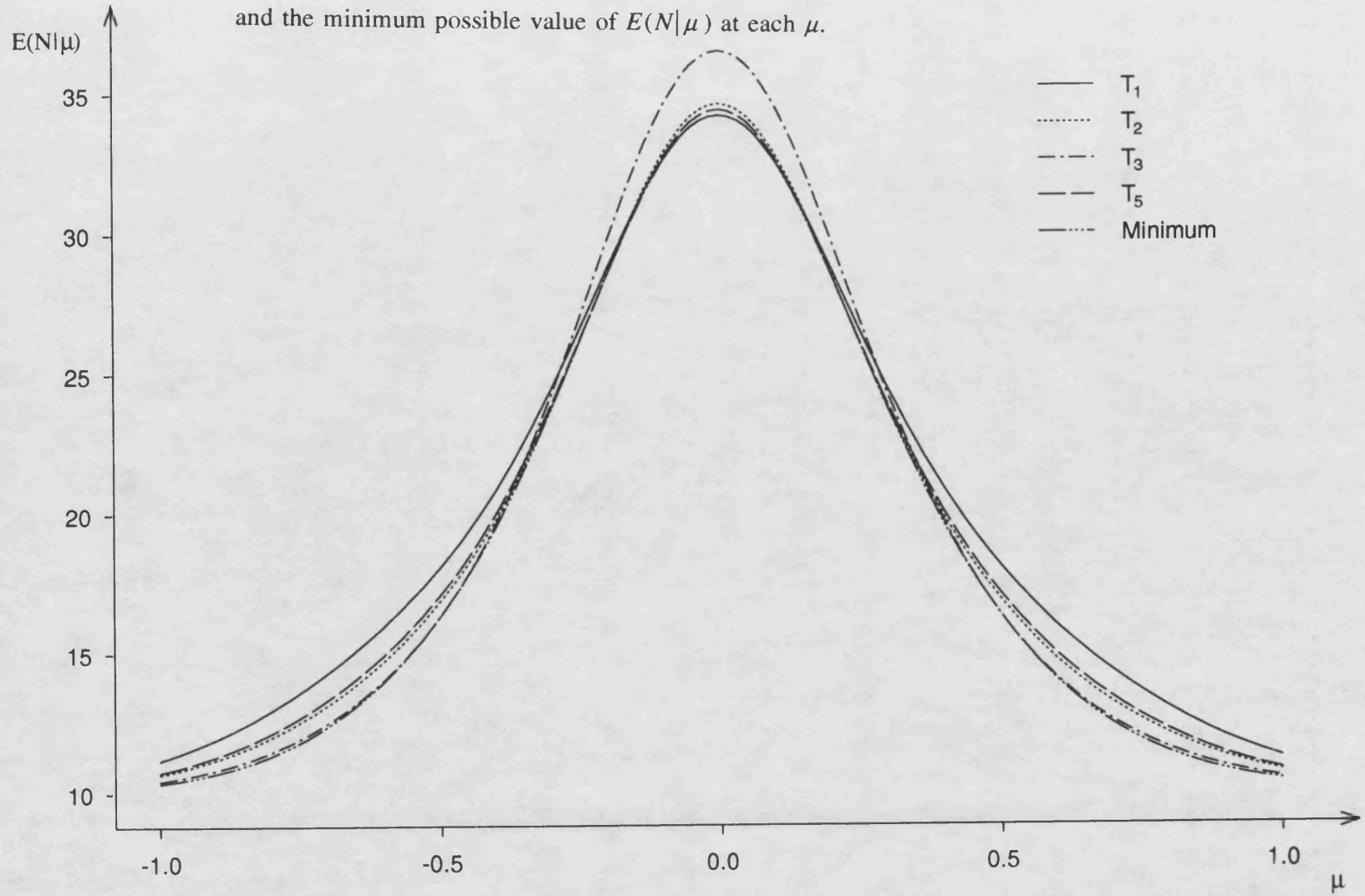
For objective function F_3 test T_3 is optimal, although none of the group sequential tests requires more than 2 observations more than T_3 on average. Test T_2 is slightly better than T_5 here, with test T_1 the poorest of the 4 sequential tests. Savings over the fixed sample size test are **at least** 27.7 observations on average here.

In terms of minimizing F_1, test T_1 is optimal. Both T_2 and T_5 are close to optimal here, with T_5 being the better of the two. Savings over the fixed sample size test are smaller here than in other cases. Even so the fixed sample size test requires 9.8 observations more than T_1 and 7.5 observations more than T_3 on average.

Figure 3.1 shows $E(N|\mu)$ plotted against μ for $-1 \leq \mu \leq 1$ (i.e. $-4\delta \leq \mu \leq 4\delta$) for T_1, T_2, T_3 and T_5. Also shown in Figure 3.1 is the curve giving the minimum of $E(N|\mu)$ for each value of μ in the range $[-1,1]$. In line with the observations made concerning Table 3.11, tests T_2 and T_5 can be seen to be optimal or close to optimal over μ in this range. Test T_1 is optimal for $\mu = 0$, close to optimal for $|\mu|$ small and sub-optimal for $|\mu|$ large. In contrast with T_1, test T_3 is sub-optimal for $|\mu|$ small and optimal or near optimal for $|\mu|$ large.

It should be noted that the four optimal tests differ slightly in their operating characteristics, that is the probability of accepting H_1^- as a function of μ (which, of course, is simply the mirror image of the probability of accepting H_1^+ as a function of μ). Symmetry and the error constraints ensure agreement at $\mu = 0$ and $\mu = \pm 0.25$, while numerical calculations show that the largest difference in operating characteristics between any two tests is less than 0.002. Hence it would seem reasonable to base comparisons between T_1, T_2, T_3 and T_5 on expected sample size alone.

Figure 3.1. Expected sample size function, $E(N|\mu)$ for tests T_1, T_2, T_3 and T_5, and the minimum possible value of $E(N|\mu)$ at each μ.



To conclude, unless there exist particularly strong *a priori* reasons for adopting tests such as T_1 or T_3, we would recommend the use of either T_2 or T_5 in practice. Clearly both T_2 and T_5 are optimal, or close to optimal over the entire parameter space for μ. Note, however, that all four tests considered here have clear advantages over the corresponding fixed sample size test.

3.12 Optimization over Unequal Group Sizes.

In §§3.5–3.11 we have considered the computation of optimal group sequential tests for problems with equally sized groups of observations. By allowing unequal group sizes we introduce an extra level of complexity into the design stage of the experiment. On the other hand, allowing group sizes to vary may lead to substantial savings in the expected sample size. In this section we consider the minimization of $F_1: E(N|\mu=0)$, $F_2: E(N|\mu=\delta)$, $F_3: E(N|\mu=2\delta)$ and $F_5: \int E(N|\mu)\delta^{-1}\varphi(\mu/\delta)\,d\mu$ over group sizes as well as feasible stopping rules.

Before describing our approach in more detail we note that this topic has received scant attention in the literature. Colton & McPherson (1976) considered the minimization of 2 group one-sided sequential tests over group sizes. They concentrated on the unsymmetric problem of §3.3 with the restriction that early stopping with the acceptance of H_0 was not permitted (i.e. $c_1' = -\infty$). They pointed out that such a restriction will not normally amount to a serious ethical problem. (Indeed in our clinical trial example of §3.2 and §3.3 a case could be made out for continuing to sample when the standard treatment is no worse than the experimental in order to assess secondary issues.) The two remaining critical values, c_1 and c_2, are constrained by the Type I and Type II error rates. By considering percentage points of the bivariate normal distribution, tests were derived which minimize $E(N|\mu=\delta)$ for both equally and unequally sized groups. Colton & McPherson pointed to the efficiency of these two stage designs with large savings in $E(N|\mu=\delta)$ over the single sample test.

For our approach to tests with unequal group sizes, consider again the problem described in §3.5, but this time with K groups of sizes n_1, $n_2 - n_1$, ..., $n_K - n_{K-1}$, where $n_1 < n_2 < \ldots < n_K$ and, in general, group sizes are not equal. For fixed K, α, δ, n_1, $n_2,\ldots,n_{K-1}$ and n_K, we can easily compute

an optimal test for a given objective function, F, by simply generalizing the method of §§3.6–3.8.

We can minimize F over group sizes as well as feasible stopping rules by searching over $\{\ln n_1, \ln(n_2 - n_1), \ldots, \ln(n_K - n_{K-1})\} \in \mathbf{R}^K$. At each stage of the search we fix group sizes and compute the optimal feasible test in the way outlined above. We use the simplex algorithm of Nelder & Mead (1965) to search over group sizes. This algorithm requires the specification of $(K+1)$ starting values when working in K dimensions, but does not require the calculation of derivatives of the objective function. As we noted in §3.5 the Nelder & Mead algorithm is only really powerful in 7 or fewer dimensions. Fortunately, for this problem, the most interesting results occur with $K \leq 5$.

Table 3.12 gives the minima of F_1, F_2, F_3 and F_5 for tests with K = 2, 3, 4, and 5, α = 0.05 and no constraints on the group sizes. Results are, again, expressed as a percentage of the corresponding fixed sample size, N_f, and are independent of δ and σ^2. The figures in parentheses are the corresponding minima for designs constrained to have equally sized groups.

Table 3.12. Minima of F_1, F_2, F_3 and F_5 for tests with $\alpha = 0.05$, K = 2, 3, 4 and 5 and no constraints on group size. Results are expressed as a percentage of the corresponding fixed sample size, N_f. The figures in parentheses are the corresponding minimum values for designs with equally sized groups.

K	F_1	F_2	F_3	F_5
2	86.5 (87.0)	71.2 (72.7)	49.3 (57.0)	76.4 (76.7)
3	81.7 (82.4)	63.9 (64.8)	38.2 (45.3)	69.4 (69.7)
4	79.2 (80.0)	60.0 (61.0)	33.6 (39.9)	66.0 (66.4)
5	77.7 (78.4)	57.7 (58.7)	31.2 (36.8)	64.2 (64.5)

Clearly, allowing unequally sized groups leads to only small percentage gains in efficiency for objective functions F_1, F_2 and F_5. Such gains could well be

outweighed by the additional complexities introduced into the design stage of the trial. Somewhat more substantial gains occur with objective function F_3. We shall consider these tests in more detail.

Table 3.13 gives the cumulative group sizes, n_1, n_2, ..., n_K, expressed as proportions of the fixed sample size, for the optimal tests for F_3 considered in Table 3.12. To ease notation we let n_i/N_f be denoted by t_i $(i = 1, 2, \ldots, K)$.

Table 3.13. The cumulative group sizes for the optimal tests for F_3 considered in Table 3.12, expressed as proportions of the fixed sample size, N_f.

	Group Sizes				
K	t_1	t_2	t_3	t_4	t_5
2	0.39	1.33			
3	0.24	0.6	1.79		
4	0.17	0.38	0.77	2.2	
5	0.15	0.29	0.51	0.95	2.65

The general pattern in Table 3.13 is for the first $K-1$ analyses to occur relatively early compared with the Kth analysis. For instance with $K = 2$ the first analysis is conducted after only 29.3% of the maximum number of patients required by the test have responded. When $K = 5$ the 4th analysis is conducted after 35.8% of the maximum number of patients have responded.

For $K \geq 3$ the maximum sample size is large compared with the corresponding fixed sample size. Indeed when $K = 4$ and $K = 5$ the maximum sample size is more than twice as large as N_f. This would tend to suggest that while these tests may be very efficient for $|\mu|$ large, they are likely to be sub-optimal when μ is small in absolute value. To confirm this we calculated $F_1: E(N|\mu=0)$, $E(N|\mu=\delta/2)$, which, to ease notation, we shall denote by $F_{\delta/2}$, and $F_2: E(N|\mu=\delta)$ for each of the 4 tests. The results, expressed as a percentage of N_f, are given in Table 3.14.

Table 3.14. Attained values of F_1, $E(N|\mu=\delta/2)$ and F_2 for the optimal tests for F_3 considered in Table 3.12 and Table 3.13.

	K			
	2	3	4	5
F_1	96.3	104.8	115.4	127.3
$F_{\delta/2}$	90.8	95.1	101.7	108.9
F_2	77.1	73.1	72.7	72.6

For $K>2$ and attained values of F_1 our tests require more patients on average than does the corresponding fixed sample size test. Moreover all 4 tests are substantially worse than the optimal tests for F_1 given in Table 3.12. Even under $\mu=\delta/2$ it would, on average, be preferable to adopt a single sample test than our optimal test with 5 groups. The attained values of F_2 are at most 77.1% of the fixed sample size. However the tests are clearly sub-optimal when attained values of F_2 are compared with the results for the optimal tests for this objective function given in Table 3.12.

To conclude, it should be said that for objective functions F_1, F_2 and F_5 allowing unequal group sizes leads to only small gains in expected sample size. While more substantial gains are obtained for objective function F_3 the resulting tests are rather sensitive to the actual treatment difference, μ. Indeed, if $\mu=0$, these tests perform worse on average than does the fixed sample size test. We would recommend the employment of tests with equally sized groups unless there are strong reasons to the contrary.

3.13 Unpredictable Numbers of Groups and Group Sizes.

In practice group sizes and the maximum number of analyses may not be known at the design stage of a study. Typically a clinical trial monitoring committee will arrange to meet at equally spaced intervals in time (every 6 months, say) over a fixed length of time. Clearly the number of patients

responding between analyses is likely to be unpredictable. The maximum number of analyses might also be unpredictable. For example, if patient recruitment on to the study is slower than anticipated the length of the trial may well be extended and the maximum number of analyses increased.

Pocock (1977) was one of the first to consider the problem of unpredictable group sizes. He conducted a simulation study into the effect of unpredictable group sizes on the error rates of his two-sided group sequential tests (described in §4.4). Using a Poisson process to generate group sizes, Pocock showed that his tests were robust to such unpredictability.

Lan & DeMets (1983) proposed a procedure flexible enough to be used when both group sizes and the maximum number of groups are unpredictable. Their approach was based on the use of an error spending function (see the discussion of Jennison's (1987) paper in §3.5). Suppose the planned maximum sample size of our test was $n_{\max}$. The function $\alpha^*(r)$ specifies the error to be "spent" after a proportion r of $n_{\max}$ has been observed. The choice of $\alpha^*(r)$ is open to the experimenter, although the function is defined to satisfy $\alpha^*(0) = 0$ and $\alpha^*(r) = \alpha$ for $r \geq 1$.

Letting $S_{n_i} = X_1 + X_2 + \ldots + X_{n_i}$, the critical values of the Lan & DeMets test are computed numerically as the solutions of the following system of equations for $i = 1, 2, \ldots, K$,

$$\sum_{j=1}^{i} \Pr_{-\delta}(|S_{n_1}| < c_1, \ldots, |S_{n_{j-1}}| < c_{j-1}, S_{n_j} > c_j) = \alpha^*(n_i/n_{\max}). \qquad (3.13.1)$$

An important feature of these equations is that c_i depends only on $n_1, n_2, \ldots, n_i$ and $c_1, c_2, \ldots, c_{i-1}$. Hence it is not necessary to know future group sizes, critical values or the maximum number of analyses in order to compute the present critical value.

The experiment is terminated either before $n_{\max}$ or at the first analysis such that the accumulated sample size is greater than $n_{\max}$. Therefore α^* should be such that $c_i = 0$ at the first analysis at which $n_i \geq n_{\max}$. In general α^* can only satisfy this last condition if the sequence $n_1, n_2, \ldots$ is known in advance. However, in practice, it suffices to choose α^* to satisfy this condition for an anticipated sequence of group sizes and then to set $c_i = 0$ at the first analysis at which $n_i \geq n_{\max}$; this will lead to only minor departures from the stated error

probabilities α at $\mu = \pm\delta$.

Lan & DeMets considered three error spending functions, namely

(i) $\alpha_1^*(r) = 2(1-\Phi(\Phi^{-1}(1-\alpha/2)/\sqrt{r}))$ $(0<r\leq 1)$

(ii) $\alpha_2^*(r) = \alpha \ln\{1+(e-1)r\}$

(iii) $\alpha_3^*(r) = \alpha r$.

The function α_1^* has very similar properties to the two-sided test of O'Brien & Fleming (1979) (described in §4.4) with a rather conservative stopping rule at early analyses. In contrast α_2^* is similar to the test of Pocock (1977) with a liberal stopping rule at early analyses. The third error spending function, α_3^*, can be seen to be a compromise between α_1^* and α_2^*.

In order to compute a group sequential test which is both flexible and highly efficient with respect to some objective function, F, we propose to use Lan & DeMets' method with the error spending function of an optimal test for F. Our only problem is that the Lan & DeMets approach requires $\alpha^*(r)$ to be continuous in r $(0\leq r\leq 1)$. We propose to use the error spending function of an optimal group sequential test with a large number, 100 say, of equally sized groups, which clearly would be discrete in r, and then to use linear interpolation to obtain a continuous function.

As an example, suppose $\sigma^2 = 2$ and $\delta = 0.2$, so that a fixed sample test with error probabilities 0.05 would require 136 observations, and it is desired to construct a group sequential test with a maximum sample size of 150 and with a low expected sample size at $\mu = \pm\delta$. We first compute the boundary values $\{\tilde{c}_1,\ldots,\tilde{c}_{100}\}$ for the optimal group sequential test for this problem with 100 groups of observations; there is no difficulty here in dealing with non-integer group sizes, we simply set $n_i = 1.5i$ and treat S_{n_i} as having a normal distribution with mean $1.5i\mu$ and variance $1.5i\sigma^2$. We then calculate

$$\alpha^*(i/100) = \sum_{j=1}^{i} \Pr_{-\delta}\{|S_{n_1}|<\tilde{c}_1,\ldots,|S_{n_{j-1}}|<\tilde{c}_{j-1}, S_{n_j}>\tilde{c}_j\} \qquad j=1,\ldots,100$$

and define the remainder of $\alpha^*(r)$ for $0\leq r\leq 1$ by linear interpolation between these values. Suppose the group sizes actually obtained are 20, 35, 40, 20, 25 and 20. Solving (3.13.1) for $i=1,\ldots,5$ and setting $c_6 = 0$ gives $c_1 = 17.0$, $c_2 = 13.9$, $c_3 = 11.2$, $c_4 = 10.7$ and $c_5 = 6.5$. Note that this test has $n_6 = 160 > n_{\max}$; however, since c_5 is so small, it is unlikely that the 6th group

of observations will be needed. This test has error probabilities 0.046 and expected sample size 88.2 at $\mu = \pm\delta$. The minimum possible $E(N|\mu=\delta)$ for a test with the same group sizes and error probabilities 0.046 at $\mu = \pm\delta$ is 87.0 , thus, the method has yielded a highly efficient test even though the actual group sizes were unknown.

In §3.5, we discussed the family of one-sided tests proposed by Emerson & Fleming (1989). The boundaries of these tests are indexed by a single parameter, Δ. For a pre-specified Δ the critical values of the Emerson & Fleming test are given by $c_i = (in)^{\Delta} z - in\delta$ $(i = 1,\ldots,K)$, with z constrained to give a test of size α and n chosen so that $c_K = 0$. The family is easily generalized to the case of fixed but unequal group sizes, n_1, $n_2 - n_1, \ldots, n_K - n_{K-1}$, by setting $c_i = n_i^{\Delta} z - n_i\delta$ $(i = 1,\ldots,K)$ with n_K constrained so that $c_K = 0$.

Emerson & Fleming went on to suggest the extension of their method to practical examples where both group sizes and the maximum number of analyses are unpredictable. At the design stage of the trial Δ and α are fixed and K is predicted. The relevant z, z', say, is calculated, and the group size n chosen so that $c_K = 0$. Suppose the actual maximum number of analyses is K' and the group sizes are n'_1, $n'_2 - n'_1, \ldots, n'_{K'} - n'_{K'-1}$, our critical values are then given by

$$c'_i = (n'_i)^{\Delta} z' - n'_i\delta \qquad i=1,2,\ldots,K'-1$$

and

$$c'_{K'} = 0.$$

Obviously K' and $n'_{K'}$ may be less than, equal to or greater than K and Kn respectively. Sampling is terminated at the first analysis after Kn observations have accrued.

Emerson & Fleming conducted a study of the attained significance levels of their method over a wide range of possible scenarios. For $\alpha = 0.05$, $\Delta = 0$ and 0.5 and K predicted to be 4, n and z' were calculated. Tests were then simulated with actual numbers of analyses K' = 2, 3, 4, 5 and 6 and group sizes

$$n'_i = \left\lceil \frac{i}{K} \right\rceil^r \pi'_K Kn$$

where r = 0.8, 1.0 and 1.5, and π'_K = 0.9, 1.0, 1.1. Obviously π'_K determines the attained maximum sample size. The parameter r determines whether group sizes are equal ($r = 1.0$), larger at early analyses ($r = 0.8$) or larger at later analyses ($r = 1.5$).

The results of Emerson & Fleming are repeated here in Tables 3.15 ($\Delta = 0$) and 3.16 ($\Delta = 0.5$). Also included is the ratio of $E(N|\mu=\delta)$ to its corresponding minimum for each of the simulated designs - this ratio is known as an "efficiency ratio".

Table 3.15. Attained Type I errors and efficiency ratios for the simulated designs considered by Emerson & Fleming (1989) for r = 0.8, 1.0 and 1.5, $\pi'_{K'}$ = 0.9, 1.0 and 1.1 and K' = 2, 3, 4, 5 and 6. The design on which these simulated tests are based has $\Delta = 0$, $K = 4$ and $\alpha = 0.05$.

		K'									
r	$\pi'_{K'}$	2		3		4		5		6	
.8	.9	.0546	1.0	.0556	.9771	.0563	.9679	.0569	.9641	.0573	.9618
	1.0	.0472	1.0	.0492	.9625	.0508	.9426	.0520	.9328	.0530	.9271
	1.1	.0421	1.0	.0457	.9486	.0484	.9164	.0504	.8997	.0518	.8909
1.0	.9	.0542	1.0	.0552	.9754	.0559	.9721	.0565	.9693	.0570	.9663
	1.0	.0464	1.0	.0484	.9565	.0500	.9458	.0512	.9398	.0522	.9343
	1.1	.0407	1.0	.0442	.9355	.0470	.9143	.0491	.9042	.0507	.8970
1.5	.9	.0535	1.0	.0545	.9949	.0552	.9820	.0558	.9770	.0562	.9735
	1.0	.0450	1.0	.0469	.9859	.0484	.9664	.0496	.9551	.0506	.9480
	1.1	.0383	1.0	.0415	.9705	.0441	.9462	.0462	.9279	.0480	.9165

Table 3.16. Attained Type I errors and efficiency ratios for the simulated designs considered by Emerson & Fleming (1989) for r = 0.8, 1.0 and 1.5, $\pi'_{K'}$ = 0.9, 1.0 and 1.1 and K' = 2, 3, 4, 5 and 6. The design on which these simulated tests are based has $\Delta = 0.5$, $K = 4$ and $\alpha = 0.05$.

		K'				
r	$\pi'_{K'}$	2	3	4	5	6
.8	.9	.0375 1.0	.0445 .9973	.0502 .9958	.0550 .9951	.0590 .9948
	1.0	.0328 1.0	.0405 .9951	.0465 .9921	.0514 .9907	.0555 .9901
	1.1	.0294 1.0	.0376 .9931	.0437 .9887	.0487 .9866	.0528 .9858
1.0	.9	.0389 1.0	.0471 .9975	.0537 .9970	.0593 .9969	.0640 .9966
	1.0	.0341 1.0	.0430 .9945	.0500 .9929	.0557 .9926	.0606 .9925
	1.1	.0307 1.0	.0401 .9913	.0473 .9881	.0531 .9876	.0580 .9879
1.5	.9	.0416 1.0	.0520 1.0	.0607 .9952	.0680 .9888	.0743 .9824
	1.0	.0366 1.0	.0478 .9994	.0569 .9958	.0644 .9902	.0709 .9842
	1.1	.0330 1.0	.0447 .9971	.0541 .9943	.0618 .9895	.0683 .9842

The attained error rates are quite impressive, ranging from 0.0383 to 0.0573 when $\Delta = 0$ and from 0.0294 to 0.0743 when $\Delta = 0.5$. When the predicted maximum sample sizes are attained (i.e. $\pi'_K = 1.0$) these errors are even more impressive, with a range of 0.045 to 0.053 when $\Delta = 0$ and 0.0328 to 0.0709 when $\Delta = 0.5$.

The efficiency ratios are also good. When $K' = 2$ these ratios are always equal to one. With $\Delta = 0$ the tabulated tests are at worst 89.09% efficient. With $\Delta = 0.5$ the tests are never worse than 98.24% efficient.

Patently the Emerson & Fleming method performs well both in terms of attained significance levels and expected sample size under $\mu = \delta$. However we claim that our proposal for dealing with both problems of unpredictability is superior. Table 3.17 gives the attained sizes and efficiency ratios under for the

simulated designs considered by Emerson & Fleming for our method based on the optimal test for F_2 with $t = 1.2$, $K = 100$ and $\alpha = 0.05$.

Table 3.17. Attained Type I errors and efficiency ratios for the simulated designs considered by Emerson & Fleming (1989) for $r = 0.8$, 1.0 and 1.5, $\pi'_{K'} = 0.9$, 1.0 and 1.1 and $K' = 2$, 3, 4, 5 and 6. The design on which these simulated tests are based is our optimal group sequential test for F_2 with $t = 1.2$, $K = 100$ and $\alpha = 0.05$.

		K'				
r	$\pi'_{K'}$	2	3	4	5	6
.8	.9	.0507 1.0	.0522 .9998	.0528 .9991	.0532 .9984	.0534 .9980
	1.0	.0452 1.0	.0471 .9990	.0479 .9982	.0483 .9978	.0486 .9977
	1.1	.0414 1.0	.0437 .9975	.0448 .9952	.0455 .9934	.0460 .9916
1.0	.9	.0507 1.0	.0521 .9980	.0527 .9974	.0531 .9976	.0586 .9978
	1.0	.0451 1.0	.0469 .9959	.0477 .9949	.0482 .9955	.0485 .9965
	1.1	.0411 1.0	.0435 .9931	.0445 .9904	.0452 .9901	.0457 .9905
1.5	.9	.0492 1.0	.0513 .9937	.0523 .9979	.0528 1.0	.0531 .9991
	1.0	.0432 1.0	.0459 .9887	.0471 .9946	.0478 .9970	.0482 .9978
	1.1	.0388 1.0	.0421 .9826	.0436 .9894	.0445 .9929	.0450 .9939

Our method produces tests with Type I errors ranging from 0.0388 to 0.0586. For $\pi'_{K'} = 1.0$ this range reduces to 0.0432 to 0.0486. These impressive results are similar to those observed for the Emerson & Fleming test with $\Delta = 0$ and superior to those in Table 3.16 when $\Delta = 0.5$. The efficiency of our tests is no less than 98.26% for the designs tabulated in Table 3.17. In terms of efficiency, then, our method is a significant improvement on the Emerson & Fleming test with $\Delta = 0$ and comparable with the results for

$\Delta = 0.5$. Other examples follow a similar pattern.

3.14 Discussion and Conclusions.

In this chapter we have introduced a computationally efficient and numerically stable method for the derivation of optimal one-sided group sequential tests on the mean of a normal distribution with known variance. Up until comparatively recently the computation of optimal tests was impractical if not impossible. Instead tests were designed which required only a set of tables and a pocket calculator for their implementation. Some of these tests had the added bonus of being close to optimal for one or more objective functions.

The increase in both the power and availability of computers means that optimal tests may now be considered. Jennison (1987) was the first to derive fully optimal one-sided group sequential tests. Our approach can be seen to be an improvement on his. In particular we can obtain optimal tests for designs with a large number of groups and use them to compute flexible procedures for the sort of problem usually encountered in practice where both the group sizes and maximum number of groups are unpredictable. Simulations show that our method compares favourably with similar techniques suggested in the literature.

In Chapters 4 and 5 we extend our method to the problems of two-sided tests with or without the option to stop early and accept the null hypothesis respectively.

Appendix.

Appendix 3.1.

Consider again the symmetric one-sided group sequential hypothesis testing problem of §3.5. For any given stopping rule we can calculate the operating characteristic function (that is the probability of accepting H_1^+ as a function of μ) and the expected sample size function, by multiple numerical integration.

For example, consider the calculation of $E(N|\mu)$. By analogy with equation (3.3.3), we have

$E(N|\mu) =$

$$n\sum_{j=1}^{K} j\int_{r_j}\int_{-c_{j-1}}^{c_{j-1}} \dots \int_{-c_1}^{c_1} f_\mu(x_1)\, f_\mu(x_2-x_1) \dots f_\mu(x_j-x_{j-1})\, dx_1 \dots dx_{j-1}\, dx_j \quad (3.A1.1)$$

where $r_j = \{(-\infty,-c_j]\cup[c_j,\infty)\}$ and $f_\mu(x)$ is a normal density with mean $n\mu$ and variance $n\sigma^2$.

Clearly $E(N|\mu)$ is the sum of an integral and $K-1$ multiple integrals. The first integral in the sum is simply

$$\int_{r_1} f_\mu(x_1)\, dx_1$$

which equals

$$1 - \Phi\left[\frac{c_1-n\mu}{\sqrt{n\sigma^2}}\right] + \Phi\left[\frac{-c_1-n\mu}{\sqrt{n\sigma^2}}\right]$$

where Φ is the c.d.f. of the standard normal distribution. The NAG library contains subroutines, S15ABF and S15ACF, for calculating $\Phi(.)$ and $1-\Phi(.)$ respectively.

The second term in the sum is proportional to

$$\int_{r_2}\int_{-c_1}^{c_1} f_\mu(x_2-x_1)\, f_\mu(x_1)\, dx_1\, dx_2. \quad (3.A1.2)$$

The integral with respect to x_1 here can be evaluated using Simpson's rule. A naive approach would be to choose the grid points of the rule to be equally spaced over the interval $[-c_1,c_1]$. However, if the total number of grid points is fixed, such an approach will lead to inaccurate calculations if the width of the interval,

$2c_1$, is "large".

A more robust approach involves placing a grid of $6N-1$ points $\{g_i: 1 \le i \le 6N-1\}$, where N is fixed, over x_1 according to the rule:
For $i = 1, N$

$$g_i = n\mu - \sqrt{n}\,\sigma\{3+4\ln(N/i)\}$$

For $i = N+1, 5N-1$

$$g_i = n\mu - \sqrt{n}\,\sigma\{3-6(i-N)/4N\}$$

and, for $i = 5N, 6N-1$,

$$g_i = n\mu + \sqrt{n}\,\sigma\{3+4\ln(i/(6N-1))\}.$$

This rule places a fixed number $(4N)$ of equally spaced points within 3 standard deviations of the mean of the distribution we are integrating. In the tails of the distribution grid points are placed increasingly far apart. To obtain a set of grid points, $\{g_{1,i1}: 1 \le i1 \le N_1\}$ to place over the interval $(-c_1, c_1)$ we first find g_i and g_{i+1} such that

$$g_i \le -c_1 < g_{i+1}$$

and g_l and g_{l+1} such that

$$g_l < c_1 \le g_{l+1}.$$

We then set $g_{1,1} = -c_1$, $g_{1,3} = g_{i+1}$, $g_{1,5} = g_{i+2}$, ..., $g_{1,N_1-2} = g_j$ and $g_{1,N_1} = c_1$. (Note that if $-c_1 < g_1$ and/or $c_1 > g_{6N-1}$ we set $g_{1,1} = g_1$ and/or $g_{1,N1} = g_{6N-1}$.) The grid points with even subscripts $\{g_{1,2i}: 1 \le i \le (N_1-1)/2\}$ are then positioned halfway between the neighbouring grid points with odd subscripts, i.e.

$$g_{1,2i} = \frac{1}{2}\{g_{1,2i-1} + g_{1,2i+1}\}.$$

The weights $\{w_{1,i1}: 1 \le i1 \le N_1\}$ for use in integral (3.A1.2) are given by

$$w_{1,1} = \frac{1}{3}(g_{1,2} - g_{1,1})$$

$$w_{1,2i} = \frac{4}{3}(g_{1,2i} - g_{1,2i-1}) \qquad i = 1, 2, \ldots, (N_1-1)/2$$

$$w_{1,2i+1} = \frac{1}{3}(g_{1,2i+2} - g_{1,2i}) \qquad i = 1, 2, \ldots, (N_1-3)/2$$

$$w_{1,N_1} = \frac{1}{3}(g_{1,N_1} - g_{1,N_1-1}).$$

Integral (3.A1.2) then becomes

$$\int_{r_2} \sum_{i1=1}^{N_1} w_{1,i1}\, f_\mu(g_{1,i1})\, f_\mu(x_2 - g_{1,i1})\, dx_2 \qquad (3.A1.3)$$

We store the terms $\{w_{1,i1} f_\mu(g_{1,i1}): 1 \le i1 \le N_1\}$ in an array $\{h_1(g_{1,i1}): 1 \le i1 \le N_1\}$ for future use. The integral (3.A1.3) is equal to

$$\sum_{i1=1}^{N_1} h_1(g_{1,i1}) \left\{ 1 - \Phi\left[\frac{c_2 - g_{1,i1} - n\mu}{\sqrt{n}\,\sigma}\right] + \Phi\left[\frac{-c_2 - g_{1,i1} - n\mu}{\sqrt{n}\,\sigma}\right] \right\}.$$

This sum is easily evaluated using the stored array $\{h_1(g_{1,i1}): 1 \le i1 \le N_1\}$ and the NAG library subroutines S15ABF and S15ACF.

The third term in the sum (3.A1.1) is proportional to the multiple integral

$$\int_{r_3} \int_{-c_2}^{c_2} \int_{-c_1}^{c_1} f_\mu(x_1)\, f_\mu(x_2 - x_1)\, f_\mu(x_3 - x_2)\, dx_1\, dx_2\, dx_3. \qquad (3.A1.4)$$

The first stage in evaluating the multiple integral (3.A1.4) is to evaluate the integral with respect to x_1. This is easily achieved by using the stored array $\{h_1(g_{1i1}): 1 \le i1 \le N_1\}$. The multiple integral (3.A1.4) equals

$$\int_{r_3} \int_{-c_2}^{c_2} \sum_{i1=1}^{N_1} h_1(g_{1,i1})\, f_\mu(x_2 - g_{1,i1})\, f_\mu(x_3 - x_2)\, dx_2\, dx_3. \qquad (3.A1.5)$$

We can now use Simpson's rule to evaluate the integral with respect to x_2. The relevant grid points and weights are obtained using the same rule as for the integral with respect to x_1. We shall denote the grid points by $\{g_{2,i2}: 1 \le i2 \le N_2\}$ and the weights by $\{w_{2,i2}: 1 \le i2 \le N_2\}$. Integral (3.A1.5) is then approximately equal to

$$\int_{r_3} \sum_{i2=1}^{N_2} \sum_{i1=1}^{N_1} h_1(g_{1,i1}) w_{2,i2} f_\mu(g_{2,i2} - g_{1,i1}) f_\mu(x_3 - g_{2,i2})\, dx_3. \qquad (3.A1.6)$$

We store the terms $\{\sum_{i1=1}^{N_1} h_1(g_{1,i1}) w_{2,i2} f_\mu(g_{2,i2} - g_{1,i1}): 1 \le i2 \le N_2\}$ in an array $\{h_2(g_{2,i2}): 1 \le i2 \le N_2\}$ for future use. It follows that (3.A1.6) is equal to

$$\sum_{i2=1}^{N_2} h_2(g_{2,i2}) \left\{ 1 - \Phi\left[\frac{c_3 - g_{2,i2} - n\mu}{\sqrt{n}\,\sigma}\right] + \Phi\left[\frac{-c_3 - g_{2,i2} - n\mu}{\sqrt{n}\,\sigma}\right] \right\}.$$

This sum is easily evaluated using the stored array $\{h_2(g_{2,i2}: 1 \le i2 \le N_2\}$ and the

NAG library subroutines S15ABF and S15ACF.

All other terms in the sum (3.A1.1) are calculated similarly. The Kth term equals

$$\sum_{i(K-1)=1}^{N_{K-1}} h_{K-1}(g_{K-1,i(K-1)})\left\{1-\Phi\left[\frac{c_K-g_{K-1,i(K-1)}-n\mu}{\sqrt{n}\sigma}\right]+\Phi\left[\frac{-c_K-g_{K-1,i(K-1)}-n\mu}{\sqrt{n}\sigma}\right]\right\}$$

where $\{h_{K-1}(g_{K-1,i(K-1)}): 1 \le i(K-1) \le N_{K-1}\}$ is a stored array with $i(K-1)$st element, $h_{K-1}(g_{K-1,i(K-1)})$ equal to

$$\sum_{i(K-2)=1}^{N_{K-2}} h_{K-2}(g_{K-2,i(K-2)})w_{K-1,i(K-1)}f_{\mu}(g_{K-1,i(K-1)}-g_{K-2,i(K-2)}).$$

Using the same techniques as those explained above we can calculate the operating characteristic function of a given one-sided group sequential test:

$$O.C.(\mu) = \Pr(\mathcal{A}_1^{+}|\mu).$$

Details are omitted.

Appendix 3.2.

Consider again the symmetric one-sided group sequential hypothesis testing problem introduced in §3.12, which has a maximum of K groups of sizes n_1, n_2-n_1, ..., n_K-n_{K-1}. (The problem of §3.5 is a special case of this problem with $n_1 = n_2-n_1 = ... = n_K-n_{K-1} = n$.) Using the same notation as in §3.12, let $X_1, X_2, \ldots, X_{n_K}$ be independent identically distributed normal random variables with unknown mean μ and known variance σ^2. We wish to test

$$H_1^{-}: \mu < 0 \quad vs \quad H_1^{+}: \mu > 0$$

with error rates

$$\Pr(\mathcal{A}_1^{+}|\mu=-\delta) = \Pr(\mathcal{A}_1^{-}|\mu=\delta) = \alpha. \qquad (3.A2.1)$$

Suppose the set of critical values $\{c_1, c_2, \ldots, c_K\}$ defines a feasible stopping rule for the above problem. By analogy with the results of §3.3, the expected sample size under μ for this problem, $E(N|\mu)$, is given by equation (3.A2.2)

$E(N|\mu) =$

$$\sum_{j=1}^{K} n_j \int_{r_j - c_{j-1}}^{c_{j-1}} \int \cdots \int_{-c_1}^{c_1} f_\mu(x_1) f_\mu(x_2 - x_1) \ldots f_\mu(x_j - x_{j-1})\, dx_1 \ldots dx_{j-1}\, dx_j \quad \text{(3.A2.2)}$$

where $r_j = \{(-\infty, -c_j] \cup [c_j, \infty)\}$ $(j = 1, 2, \ldots, K)$, $f_\mu(x_1)$ is a normal density with mean $n_1\mu$ and variance $n_1\sigma^2$ and $f_\mu(x_i - x_{i-1})$ $(i = 2, 3, \ldots, K)$ are normal densities with means $(n_i - n_{i-1})\mu$ and variances $(n_i - n_{i-1})\sigma^2$.

Further, the operating characteristic function given μ for this problem, $OC(\mu) = \Pr(\mathcal{A}_1^+ | \mu)$, is given by equation (3.A2.3),

$OC(\mu) =$

$$\sum_{j=1}^{K} \int_{c_j - c_{j-1}}^{\infty} \int \cdots \int_{-c_1}^{c_1} f_\mu(x_1)\, f_\mu(x_2 - x_1) \ldots f_\mu(x_j - x_{j-1})\, dx_1 \ldots dx_{j-1}\, dx_j. \quad \text{(3.A2.3)}$$

We now consider a new problem in which the variance of the original problem, σ^2, is replaced by ${\sigma_1}^2$, and the reference improvement, δ, is replaced by δ_1 (> 0). The Type I error rate of this new test, α, and the maximum number of analyses, K, remain the same as for the original problem.

Consider the following stopping rule for our new problem defined by the rescaled set of critical values $\{c_1^*, c_2^*, \ldots, c_K^*\}$, where $c_i^* = \dfrac{{\sigma_1}^2 \delta}{\sigma^2 \delta_1} c_i$ $(i = 1, 2, \ldots, K)$, together with the rescaled cumulative group sizes $n_i^* = \dfrac{{\sigma_1}^2 \delta^2}{\sigma^2 {\delta_1}^2} n_i$ $(i = 1, 2, \ldots, K)$. In this Appendix we prove the following Lemmas and Corollaries:

Lemma 3.1 : The operating characteristic function for our new problem, $OC^*(\mu)$, is such that

$$OC^*(\mu) = OC(\mu\delta/\delta_1).$$

Corollary 3.1 : From **Lemma 3.1** it follows that

$$OC^*(\delta_1) = OC(\delta)$$

and

$$OC^*(-\delta_1) = OC(-\delta).$$

As $OC(\delta) = 1-\alpha$ and $OC(-\delta) = \alpha$, it follows that the rescaled set of critical values $\{c_1^*, c_2^*, \ldots, c_K^*\}$ together with the rescaled cumulative group sizes, $n_1^*, n_2^*, \ldots, n_K^*$, define a feasible stopping rule for our new problem.

Lemma 3.2 : Letting $E(N^*|\mu)$ denote the expected sample size function for our new problem and N_f^* denote the corresponding fixed sample size, we prove that

$$\frac{E(N^*|\mu)}{N_f^*} = \frac{E(N|\mu\delta/\delta_1)}{N_f}$$

where $E(N|\mu)$ is the expected sample size function for the original problem and N_f is the corresponding fixed sample size.

Corollary 3.2 : From **Lemma 3.2** it follows that

$$(i)\quad \frac{E(N^*|\mu=0)}{N_f^*} = \frac{E(N|\mu=0)}{N_f}$$

$$(ii)\quad \frac{E(N^*|\mu=\delta_1)}{N_f^*} = \frac{E(N|\mu=\delta)}{N_f}$$

$$(iii)\quad \frac{E(N^*|\mu=2\delta_1)}{N_f^*} = \frac{E(N|\mu=2\delta)}{N_f}$$

$$(iv)\quad \frac{\int E(N^*|\mu)\,\delta_1^{-1}\,\varphi(\mu/\delta)d\mu}{N_f^*} = \frac{\int E(N|\mu')\,\delta^{-1}\,\varphi(\mu'/\delta)d\mu'}{N_f}$$

where $\mu' = \mu\delta/\delta_1$.

Proof of Lemma 3.1 :

The operating characteristic function for the new problem is given by

$OC^*(\mu) =$

$$\sum_{j=1}^{K} \int_{c_j^*}^{\infty} \int_{-c_{j-1}^*}^{c_{j-1}^*} \cdots \int_{-c_1^*}^{c_1^*} f_\mu(x_1)\, f_\mu(x_2-x_1) \ldots f_\mu(x_j-x_{j-1})\, dx_1 \ldots dx_{j-1}\, dx_j \quad (3.A2.4)$$

where $r_i^* = \{(-c_i^*, c_i^*)\}$ for $i = 1, 2, \ldots, K-1$, $f_\mu(x_1)$ is a normal density with mean $n_1^* \mu$ and variance $n_1^* \sigma^2$, and $f_\mu(x_i - x_{i-1})$ $(i = 2, 3, \ldots, K)$ are normal densities with means $(n_i^* - n_{i-1}^*)\mu$ and variances $(n_i^* - n_{i-1}^*)\sigma_1^2$.

Consider the substitutions $z_i = \dfrac{\sigma^2 \delta_1}{\sigma_1^2 \delta} x_i$ for $i = 1, 2, \ldots, K$, in equation (3.A2.4). We obtain

$OC^*(\mu) =$

$$\sum_{j=1}^{K} \int_{c_j - c_{j-1}}^{\infty} \int \cdots \int_{-c_1}^{c_1} g_{\mu'}(z_1)\, g_{\mu'}(z_2 - z_1) \cdots g_{\mu'}(z_j - z_{j-1})\, dz_1\, dz_2 \cdots dz_{j-1}\, dz_j \quad \text{(3.A2.5)}$$

where $\mu' = \mu\delta/\delta_1$, $g_{\mu'}(z_1)$ is a normal density with mean $n_1 \mu'$ and variance $n_1 \sigma^2$, $g_{\mu'}(z_i - z_{i-1})$ $(i = 2, 3, \ldots, K)$ are normal densities with means $(n_i - n_{i-1})\mu'$ and variances $(n_i - n_{i-1})\sigma^2$, and $\{c_1, c_2, \ldots, c_K\}$ is a feasible set of critical values for our original problem.

Comparing the RHS of equation (3.A2.5) with equation (3.A2.3), we see that

$$OC^*(\mu) = OC(\mu') = OC(\mu\delta/\delta_1). \quad \text{(3.A2.6)}$$

Q.E.D.

Proof of Corollary 3.1 :

Substituting $\mu = \delta_1$ in to equation (3.A2.6) gives

$$OC^*(\delta_1) = OC(\delta),$$

and from equation (3.A2.1) we have $OC(\delta) = 1 - \alpha$. Therefore $OC^*(\delta_1) = 1 - \alpha$.

Substituting $\mu = -\delta_1$ in to equation (3.A2.6) gives

$$OC^*(-\delta_1) = OC(-\delta)$$

and from equation (3.A2.1) we have $OC(-\delta) = \alpha$. Therefore $OC^*(-\delta_1) = \alpha$.

Hence the set of rescaled critical values $\{c_1^*, c_2^*, \ldots, c_K^*\}$ together with the rescaled group sizes, n_1^*, $n_2^* - n_1^*$, ..., $n_K^* - n_{K-1}^*$, define a feasible stopping rule for our new problem.

Proof of Lemma 3.2 :

The expected sample size function for our new problem is given by

$E(N^*|\mu) =$

$$\sum_{j=1}^{K} n_j^* \int_{r_j^*} \int_{-c_{j-1}^*}^{c_{j-1}^*} \cdots \int_{-c_1^*}^{c_1^*} f_\mu(x_1) f_\mu(x_2-x_1) \ldots f_\mu(x_j-x_{j-1})\, dx_1 \ldots dx_{j-1} dx_j \quad (3.A2.7)$$

where $r_j^* = \{(-\infty,-c_j^*),(c_j^*,\infty)\}$ for $j=1,2,\ldots,K$, $f_\mu(x_1)$ is a normal density with mean $n_1^*\mu$ and variance $n_1^*\sigma_1^2$, and $f_\mu(x_i-x_{i-1})$ $(i=2,\ldots,K)$ are normal densities with means $(n_i^*-n_{i-1}^*)\mu$ and variances $(n_i^*-n_{i-1}^*)\sigma^2$.

Consider the substitutions $z_i = \dfrac{\sigma^2\delta_1}{\sigma_1^2\delta} x_i$, for $i=1,2,\ldots,K$, in equation (3.A2.7). We obtain

$E(N^*|\mu) =$

$$\sum_{j=1}^{K} n_j^* \int_{r_j} \int_{-c_{j-1}}^{c_{j-1}} \cdots \int_{-c_1}^{c_1} g_{\mu'}(z_1)\, g_{\mu'}(z_2-z_1) \ldots g_{\mu'}(z_j-z_{j-1})\, dz_1 \ldots dz_{j-1} dz_j \quad (3.A2.8)$$

where $r_j = \{(-\infty,-c_j)\cup(c_j,\infty)\}$ for $j=1,2,\ldots,K$, $g_{\mu'}(z_1)$ is a normal density with mean $n_1\mu'$ and variance $n_1\sigma^2$, $g_{\mu'}(z_i-z_{i-1})$ $(i=2,3,\ldots,K)$ are normal densities with means $(n_i-n_{i-1})\sigma^2$ and variances $(n_i-n_{i-1})\sigma^2$, and $\{c_1,\ldots,c_{K-1},c_K\}$ is a feasible set of critical values for our original problem.

Comparing the RHS of equation (3.A2.8) with equation (3.A2.2), we have

$$E(N^*|\mu) = \frac{n}{n^*} E(N|\mu') \quad (3.A2.9)$$

where $E(N|\mu')$ is the expected sample size function of our original problem under a treatment difference of μ'.

From earlier we have

$$n^* = \frac{\sigma_1^2\delta^2}{\sigma^2\delta_1^2} n.$$

Substituting for n^* in equation (3.A2.9) and rearranging, we obtain

$$\frac{E(N^*|\mu)}{\sigma_1^2\delta^2} = \frac{E(N|\mu')}{\sigma^2\delta_1^2} \quad (3.A2.10)$$

and dividing both sides of equation (3.A2.10) by $\{\Phi^{-1}(1-\alpha)\}^2$ and rearranging gives

$$\frac{E(N^*|\mu)}{N_f^*} = \frac{E(N|\mu')}{N_f} \qquad (3.A2.11)$$

where, by analogy with equation (3.5.3),

$$N_f^* = \frac{{\sigma_1}^2}{{\delta_1}^2}\{\Phi^{-1}(1-\alpha)\}^2.$$

Q.E.D.

Proof of Corollary 3.2 :

(i) Substituting $\mu = 0$ in to equation (3.A2.11) gives

$$\frac{E(N^*|\mu=0)}{N_f^*} = \frac{E(N|\mu=0)}{N_f}$$

(ii) Substituting $\mu = \delta_1$ in to equation (3.A2.11) gives

$$\frac{E(N^*|\mu=\delta_1)}{N_f^*} = \frac{E(N|\mu=\delta)}{N_f}$$

(iii) Substituting $\mu = 2\delta_1$ in to equation (3.A2.11) gives

$$\frac{E(N^*|\mu=2\delta_1)}{N_f^*} = \frac{E(N|\mu=2\delta)}{N_f}$$

(iv) Consider substituting $\mu' = \delta\mu/\delta_1$ in the expression

$$\int \frac{E(N^*|\mu)}{N_f^*}\,{\delta_1}^{-1}\,\varphi(\mu/\delta_1)\,d\mu$$

we obtain

$$\int \frac{E(N^*|\mu'\delta_1/\delta)}{N_f^*}\,\delta^{-1}\,\varphi(\mu'/\delta)\,d\mu'.$$

From Lemma 3.2, it follows that,

$$\frac{\int E(N^*|\mu)\,{\delta_1}^{-1}\,\varphi(\mu/\delta_1)\,d\mu}{N_f^*} = \frac{\int E(N|\mu')\,\delta^{-1}\,\varphi(\mu'/\delta)\,d\mu'}{N_f}.$$

4. Optimal Two-Sided Group Sequential Tests.

4.1 Introduction.

In §3 we introduced a numerically stable and computationally efficient method for deriving optimal one-sided group sequential tests on the mean of a normal distribution with known variance. In both §4 and §5 we generalize our method to the problem of two-sided group sequential tests.

In §4 we consider tests which allow an experiment to be stopped early with the rejection of the null hypothesis, H_0, but acceptance of H_0 is only permitted at the final analysis. In §5 we go on to consider tests which allow early stopping for both the acceptance and rejection of H_0.

As in §3 the motivation for this work comes from a clinical trials problem. However both the designs and the results of §4 have much wider applications.

4.2 The Fixed Sample Size Test.

Consider a clinical trial with $2N$ patients available for testing the relative efficacies of two experimental treatments which we shall denote by A and B. On entry to the trial each patient is randomly assigned to one of the two treatments. An allocation method such as the randomized permuted block design is used in order to ensure that exactly N patients are entered on to each treatment arm. Let the random variable X_i $(i = 1,\ldots,N)$ represent the difference in response between the ith patient on treatment A and the ith patient on treatment B. Suppose that the X_i's are independent and normally distributed with unknown mean μ and known variance σ^2. The parameter μ is, then, a measure of treatment difference about which we would like to make inferences. For example consider the following single sample two-sided hypothesis test on μ:
We wish to test

$$H_1^-\colon \mu < 0 \qquad vs \qquad H_0\colon \mu = 0 \qquad vs \qquad H_1^+\colon \mu > 0$$

with Type I error rate

$$\Pr(\mathcal{A}_1^- \cup \mathcal{A}_1^+ \mid \mu = 0) = \alpha \tag{4.2.1}$$

where $\mathcal{A}_1^-$ and $\mathcal{A}_1^+$ denote the acceptance of H_1^- and H_1^+ respectively.

The uniformly most powerful unbiased test for this problem accepts H_0 if

$$|S_N| = |\sum_{i=1}^{N} X_i| < \sqrt{N\sigma^2}\,\Phi^{-1}(1-\alpha/2)$$

it rejects H_0 in favour of H_1^+ if

$$S_N \geq \sqrt{N\sigma^2}\,\Phi^{-1}(1-\alpha/2)$$

and it rejects H_0 in favour of H_1^- if

$$S_N \leq -\sqrt{N\sigma^2}\,\Phi^{-1}(1-\alpha/2).$$

The Type II error rates of the above test depend on the sample size. Letting $\mathcal{A}_0$ denote the acceptance of H_0, the test satisfies the Type II error constraints

$$\Pr(\mathcal{A}_0 \cup \mathcal{A}_1^+ | \mu = -\delta) = \beta \tag{4.2.2}$$

and

$$\Pr(\mathcal{A}_1^- \cup \mathcal{A}_0 | \mu = \delta) = \beta \tag{4.2.3}$$

if N is chosen equal to N_f where

$$N_f = \frac{\sigma^2}{\delta^2}\{\Phi^{-1}(1-\beta) + \Phi^{-1}(1-\alpha/2)\}^2. \tag{4.2.4}$$

The above fixed sample size test is widely used in clinical trials. However, as was mentioned in §3.2, there often exist strong economic and ethical arguments for the adoption of group sequential rather than the fixed sample size tests. In §4.3 we outline a two-sided group sequential test on the mean of a normal distribution with known variance.

4.3 Two-Sided Group Sequential Tests.

Consider again the two-sided hypothesis testing problem outlined at the start of §4.2. This time, however, a maximum of K groups of $2n$ patients are available for entry on to the trial with n patients in each group being randomly assigned to treatment A and the remaining n to treatment B. At the ith analysis ($i = 1, \ldots, K-1$) we can either stop the trial and reject H_0 in favour of H_1^- or H_1^+, or we can admit the next group of $2n$ patients on to the experiment. If the trial continues to the Kth analysis, it is terminated with either the acceptance or the rejection of H_0.

As was mentioned in §4.1, this test does not permit a trial to be stopped early with the acceptance of H_0. This does not necessarily constitute a serious ethical problem as early evidence in favour of H_0 would suggest that the treatments are equally effective. Indeed continuing the trial could be advantageous as it would allow the experimenter to assess secondary issues such as the side effects of the treatments.

So we shall consider symmetric stopping rules of the general form :-

At analysis i $(1 \le i \le K-1)$,

if $S_{in} \ge c_i$ stop entering patients on to the trial and accept H_1^+;

if $S_{in} \le -c_i$ stop entering patients on to the trial and accept H_1^-;

if $|S_{in}| < c_i$ enter the next group of $2n$ patients on to the trial.

At analysis K,

if $S_{Kn} \ge c_K$ stop entering patients on to the trial and accept H_1^+;

if $S_{Kn} \le -c_K$ stop entering patients on to the trial and accept H_1^-;

if $|S_{Kn}| < c_K$ stop entering patients on to the trial and accept H_0.

Given n and K, we can use numerical methods to compute a set of critical values satisfying the error constraints (4.2.1)-(4.2.3). We term such a set **feasible**, and a test with a feasible stopping rule a **feasible test**. From equations (4.2.1)-(4.2.3) we require $c_1, c_2, \ldots, c_K$ to satisfy

$$\sum_{j=1}^{K} \Pr\left(|S_n| < c_1, \ldots, |S_{(j-1)n}| < c_{j-1}, |S_{jn}| \ge c_j \mid \mu = 0\right) = \alpha \qquad (4.3.1)$$

$$\sum_{j=1}^{K} \Pr\left(|S_n| < c_1, \ldots, |S_{(j-1)n}| < c_{j-1}, S_{jn} \in r_j^{(1)} \mid \mu = \delta\right) = \beta \qquad (4.3.2)$$

where $r_j^{(1)} = (-\infty, -c_j)$ for $j < K$ and $r_K^{(1)} = (-\infty, c_K)$, and

$$\sum_{j=1}^{K} \Pr\left(|S_n| < c_1, \ldots, |S_{(j-1)n}| < c_{j-1}, S_{jn} \in r_j^{(2)} \mid \mu = -\delta\right) = \beta \qquad (4.3.3)$$

where $r_j^{(2)} = (c_j, \infty)$ for $j < K$ and $r_K^{(2)} = (-c_K, \infty)$.

These joint probabilities may be expressed in terms of multiple integrals and evaluated by multiple numerical integration based on Simpson's rule (see Appendix 5.1 for further details).

Our group sequential test may also be considered in terms of its nominal significance levels $\alpha_1, \alpha_2, \ldots, \alpha_K$. The ith nominal significance level, α_i, is simply the Type I error of the fixed sample size test with in observations and critical values $\pm c_i$, i.e.

$$\alpha_i = \Pr(|S_{in}| \geq c_i \,|\, \mu = 0) = 2\left\{1 - \Phi\left[\frac{c_i}{\sqrt{in\sigma^2}}\right]\right\}.$$

So in the earlier discussion the critical value c_i may be replaced by $\sqrt{in\sigma^2}\Phi^{-1}(1-\alpha_i/2)$.

Clearly the sample size of our two-sided group sequential test is not fixed *a priori*, but is a random variable. The expected number of patients on each treatment arm under a given treatment difference μ, $E(N|\mu)$, is given by equation (4.3.4)

$$E(N|\mu) = n\sum_{j=1}^{K} j \Pr(|S_n| < c_1, \ldots, |S_{n(j-1)}| < c_{j-1}, |S_{nj}| \in r_j^{(3)} |\, \mu) \quad (4.3.4)$$

where $r_j^{(3)} = \{(-\infty,-c_j)\cup(c_j,\infty)\}$ for $j<K$ and $r_K^{(3)} = \mathbf{R}$.

Two-sided group sequential tests with the same error rates may be compared in terms of their expected and maximum sample sizes. Clearly the maximum sample size of the above sequential test is $2nK$, while the maximum number of patients on each treatment arm is nK. A comparison of two-sided group sequential tests in terms of their maximum and expected sample sizes is given in §4.9.

4.4 A Review of the Literature on Two-Sided Sequential Tests.

There is a substantial body of literature relating to two-sided sequential tests. As with the one-sided sequential tests discussed in §3.4, the earliest two-sided tests were fully sequential (i.e. $n = 1$).

Fully Sequential Two-Sided Tests.

Armitage (1957) suggested setting $c_i = a + bi$ $(i = 1, 2, \ldots, K)$ where a and b are chosen to satisfy the error constraints (4.2.1)-(4.2.3). To obtain good approximations for a and b Armitage considered an analogous problem in

continuous time. The feasible stopping rule for the continuous time problem is readily obtained and this can be used to give an approximately feasible stopping rule for the original discrete problem. We note that with a modern computer the relevant values of a and b could be obtained directly.

Armitage, McPherson & Rowe (1969) and McPherson (1974) considered tests with constant nominal significance levels (i.e. $\alpha_1 = \alpha_2 = \ldots = \alpha_K = \alpha'$, say). They showed that for given α' the overall Type I error of a test, α, increases with K and that, even for moderately large K, can be much greater than the nominal level, α'. For example with $K=10$ and $\alpha'=0.05$ the probability of wrongly rejecting H_0 is 0.19. Indeed, as was pointed out by Robbins (1952), $\alpha \to 1$ as $K \to \infty$ by the law of the iterated logarithm.

In order to preserve the overall size of the test, Armitage *et al.* (1969) suggested conducting interim tests at a very stringent nominal significance level. For given α and K the relevant α' was computed by numerical methods. Table 4.1 gives α' for $\alpha = 0.05$ and $K = 1, 5, 10, 20, 50, 100$ and 200.

Table 4.1. The nominal significance level α' leading to a test with overall significance level $\alpha=0.05$ for the Armitage, McPherson & Rowe (1969) test with $K = 1, 5, 10, 20, 50, 100$ and 200.

K	1	5	10	20	50	100	200
α'	0.05	0.015	0.01	0.007	0.005	0.004	0.003

It is important to realize that the results given in Table 4.1 are independent of σ^2 and δ. Clearly the effect of increasing K is to decrease α' which in turn increases the critical values of the test given by $c_i = \sqrt{in\sigma^2}\,\Phi^{-1}(1-\alpha'/2)$.

Group Sequential Two-Sided Tests.

Pocock (1977) pointed out that fully sequential methods are likely to be impractical in the context of a clinical trial. He proposed a two-sided group sequential test (i.e. $n>1$) with constant nominal significance levels. For a given problem with K and α fixed, the relevant nominal level is independent of group

size. Hence for $\alpha=0.05$ and K = 1, 5, 10, 20, 50, 100 and 200 the nominal significance levels for Pocock's test are given in Table 4.1. For tests with other values of α and K, nominal levels can be computed using the same techniques as Armitage *et al.*. The group size, n, is then chosen to satisfy the Type II error constraints (4.2.2) and (4.2.3).

Pocock's test is relatively simple to implement and quite insensitive to small changes in group sizes. The test has the added advantage of being close to optimal in terms of minimizing $E(N|\mu=\delta)$ and $E(N|\mu=-\delta)$ over the set of feasible tests for β "small". Furthermore, for reasonably large group sizes, the test is robust to departures from normality (by the Central Limit Theorem) and to the assumption that the population variance, σ^2, is known.

There are, however, strong arguments for considering tests with increasing nominal significance levels (i.e. $\alpha_1 < \alpha_2 < \ldots < \alpha_K$). For instance Pocock's test is often criticized for being too liberal at early analyses when the accumulated sample size tends to be small. Many people have argued that a rather extreme test statistic would be required at, say, the first interim analysis for the trial's monitoring committee to want to stop the experiment and reject H_0. One possible reason for this could be the desire to continue the trial in order to assess secondary issues such as treatment side-effects.

Both Pocock (1982) and DeMets (1987) have argued for group sequential tests which resemble the corresponding fixed sample size test at the final analysis. The problem with the constant nominal significance level approach is that it requires a relatively conservative test at the final analysis. For example with $K=5$ and $\alpha=0.05$, Pocock's test would only reject H_0 at the 5th analysis if $|T_5| = |S_{5n}/\sqrt{5n\sigma^2}| \geq 2.42$. The corresponding single sample test would reject H_0 if $|S_{5n}/\sqrt{5n\sigma^2}| \geq 1.96$. So, for instance, an observed test statistic of $T_5 = 2.35$ would lead to radically different conclusions depending on how many times the data had been analysed and how much Type I error was left to be spent at the 5th analysis. This contradiction leads to confusion and lends weight to the criticisms by non-frequentist statisticians.

Several tests have been proposed in the literature which avoid the problems associated with the constant nominal significance level approach. Haybittle (1971) and Peto *et al.* (1976) suggested adopting very conservative critical values at the first $K-1$ analyses. The stopping rule for these tests set

$c_i = 3\sqrt{in\sigma^2}$ (or $\alpha_i' = 0.001$) for $i = 1, 2, \ldots, K-1$. At the Kth analysis we can either set $c_K = \Phi^{-1}(1-\alpha/2)\sqrt{Kn\sigma^2}$ and obtain a test with Type I error marginally greater than α, or we can compute the critical value c_K giving a test of size α exactly. As DeMets (1987) pointed out, the Haybittle/Peto test is extremely flexible; for instance, it is not necessary to specify the maximum number of analyses, K, at the design stage of the trial, although, as has been noted by Lan & DeMets (1989), it is only legitimate to alter K for reasons unrelated to the inferred treatment difference. The main disadvantage of the Haybittle/Peto test is that the probability of stopping early is, not surprisingly, relatively small.

O'Brien & Fleming (1979) proposed a test with critical values given by $c_i = z'$ $(i = 1, 2, \ldots, K)$ where z' is a constant chosen so that the overall size of the test equals α. Again an appropriate choice of group size leads to the Type II error constraints being satisfied. The test is rather conservative at early analyses (sometimes more conservative than even the Haybittle/Peto test!) but becomes increasingly more liberal at each stage. At the Kth analysis the O'Brien & Fleming test is similar to the fixed sample size test of size α based on the same number of observations. For example with $K = 5$ and $\alpha = 0.05$ the nominal significance levels of the O'Brien & Fleming test are given by: $\alpha_1 = 10^{-5}$, $\alpha_2 = 0.0013$, $\alpha_3 = 0.0084$, $\alpha_4 = 0.0025$ and $\alpha_5 = 0.041$. The O'Brien & Fleming test avoids many of the problems associated with the Pocock test while, in general, it offers a higher probability of stopping the trial early for $|\mu|$ large than does the Haybittle/Peto test. DeMets (1987) concluded that the O'Brien & Fleming test provides a good compromise between the earlier two approaches.

Fleming, Harrington and O'Brien (1984) defined a family of two-sided group sequential tests in terms of the Type I error spent at each analysis. Let π_i denote the Type I error spent at analysis i, then

$$\pi_i = \Pr\left(|S_n| < c_1, \ldots, |S_{(i-1)n}| < c_{i-1}, |S_{in}| \geq c_i \mid \mu = 0\right) \quad i = 1, \ldots, K.$$

For a test of size α we require that $\pi_1 + \ldots + \pi_K = \alpha$. Fleming *et al.* suggested choosing $\pi_1 = \pi_2 = \ldots = \pi_{K-1} = \pi$, say, with $\pi_K = \alpha - (K-1)\pi$. For a given choice of π, K and α numerical methods give the resulting set of critical values.

In choosing π, Fleming *et al.* recommended considering the parameter $\gamma = (K-1)\,\pi / \alpha$ which is the ratio of the Type I error spent at the first $K-1$ analyses to α. Choosing γ small (large) gives a test which is conservative (liberal)

at early analyses. Fleming *et al.* considered γ = 0.1, 0.2, ..., 0.5. A suitable choice of γ gives rise to a test which provides a compromise between the extreme conservatism typical of the early analyses of the O'Brien & Fleming and Haybittle/Peto tests and the controversial liberalism at the early analyses of the Pocock test.

Like the Haybittle/Peto test, the Fleming *et al.* test is very flexible. For instance it allows us to consider tests with unequal group sizes. For the problem outlined at the start of §4.3, but with K groups of sizes n_1, n_2-n_1, ..., n_K-n_{K-1}, this is achieved by computing π_i from the equation

$$\pi_i = \Pr\left(|S_{n_1}| < c_1, \ldots, |S_{n_{i-1}}| < c_{i-1}, |S_{n_i}| \geq c_i \mid \mu = 0\right) \quad i = 1,\ldots,K.$$

and then proceeding as for the case of equally sized groups

The Fleming *et al.* test also allows for the maximum number of analyses, K, to be altered during the course of a trial. This is a major advantage if, for example, originally a maximum of K analyses are planned at fixed points in time (e.g. every six months) but patient response turns out to be slower than anticipated. It should be noted however that, as with the Haybittle/Peto test, K can only be changed for reasons unrelated to the inferred treatment difference at interim analyses.

4.5 Optimal Two-Sided Group Sequential Tests.

The definition of an optimal two-sided group sequential test is similar to that for a one-sided group sequential test given in §3.5. For the problem outlined in §4.3 with a maximum of K groups of n observations and some given objective function, F, the optimal test minimizes F subject to the error constraints (4.2.1)-(4.2.3).

In the following discussion we shall concentrate on the three objective functions

$$F_2 = E(N|\mu=\delta)$$

$$F_3 = E(N|\mu=2\delta)$$

$$F_5 = \int E(N|\mu)\delta^{-1}\varphi(\mu/\delta)\,d\mu.$$

There is a substantial body of literature concerning the computation of optimal two-sided tests. For instance Pocock (1982) considered the minimization of F_2: $E(N|\mu=\delta)$. For fixed K and α he searched over sets of nominal significance levels for the required minimum. At each stage of the search the group size, n, was constrained to give a test with Type II error β at $\mu=\pm\delta$. Pocock demonstrated that his constant nominal significance level test is near optimal when β is small (i.e. when the power of detecting the treatment difference $\mu=\delta$ is large). He also showed that when β is large (≥ 0.5, for example) the O'Brien & Fleming (1979) test is near optimal.

Of course Pocock's observations only apply to objective function F_2. For other objective functions Pocock's test is sub-optimal when β is small.

Elashoff & Reedy (1984) considered two-group sequential tests with both equal and unequal group sizes. They were particularly interested in the effect of varying the first nominal significance level, α_1. For the tests of Pocock (1977) and O'Brien & Fleming (1979), α_1 equals 0.0294 and 0.005 respectively when $K=2$. Elashoff & Reedy considered the performance of these two tests in terms of $E(N|\mu)/N_f$ for both $\beta=0.1$ and $\beta=0.5$. In both cases the O'Brien & Fleming test is superior when $|\mu|$ is either particularly small or particularly large. For intermediate values of $|\mu|$ the Pocock test is superior. Elashoff & Reedy suggested setting $\alpha_1=0.015$ for a test which performs fairly well for all μ.

Wang & Tsiatis (1987) considered minimizing F_2 over a family of stopping rules indexed by a single parameter Δ. The critical values of these stopping rules are given by: $c_i = i^{\Delta} z$ $(i=1,\ldots,K)$, where z is constrained to give a test of size α. Two special cases of the Wang & Tsiatis family of tests are the Pocock test ($\Delta=0.5$) and the O'Brien & Fleming test ($\Delta=0$). For fixed K and α, Wang & Tsiatis searched over Δ for the "minimum" of F_2. At each stage of the search the group size was fixed to give a test with Type II error β at $\mu=\pm\delta$. By comparing their results with those of Pocock (1982), Wang & Tsiatis demonstrated that their tests are very close to being optimal. In agreement with Pocock's results they showed that the constant nominal significance level test is indeed near optimal for $\beta<0.1$, while the O'Brien & Fleming test is near optimal for $\beta>0.4$. For intermediate values of β the Wang & Tsiatis boundaries with Δ somewhere between 0 and 0.5 are approximately optimal.

Although Wang & Tsiatis only considered the single objective function, F_2, their approach could be easily extended to compute near optimal tests for other objective functions.

Clearly the problem outlined in §4.3 is more complicated than either of those considered by Pocock (1982) or Wang & Tsiatis (1987). Given n and K our problem is one of determining a set of critical values $\{c_1, c_2, \ldots, c_K\}$ which minimizes an objective function, F, subject to the error constraints (4.2.1)-(4.2.3). An obvious way of proceeding is to use a method similar to that of Jennison (1987) for one-sided tests. As the constraints (4.2.2) and (4.2.3) are effectively the same our problem has $K-2$ degrees of freedom and so we could search over $c_1, c_2, \ldots, c_{K-2}$ for the minimum of F while, at each stage of the search, constraining c_{K-1} and c_K to give a feasible test. The pair c_{K-1} and c_K are obtained by using the method of Powell (1970) for solving two non-linear simultaneous equations in two unknowns. (Powell's algorithm is available as a subroutine of the NAG library.) When no such pair exists the objective function would be assigned a high positive value to move the search away from infeasible regions.

With only two analyses ($K = 2$) our problem is simply to obtain the unique pair $\{c_1, c_2\}$ giving rise to a feasible test. When K=3 we are faced with a one-dimensional minimization problem and we use the Golden Section Search algorithm. When K is greater than 3 we have a multi-dimensional minimization problem and we use the simplex algorithm of Nelder & Mead (1965).

It is not difficult to see that the above approach suffers from the same inadequacies as Jennison's method. Slow convergence, no guarantee of convergence to the global minimum and unreliability in more than about 7 dimensions force us to consider an alternative approach.

In §4.6 we propose an improved method for the computation of optimal two-sided group sequential tests. Our improved method has the important advantages of being computationally efficient and numerically stable.

The approach is a generalization of the method for computing optimal one-sided group sequential tests described in §3.6. Initially we consider a family of problems in Bayesian decision theory with a common prior distribution and cost of sampling function. The forms of the prior and cost function are determined by the objective function we are interested in minimizing. Individual problems within

the family differ in terms of their loss functions. These loss functions are indexed by a pair of loss parameters, d_0 and d_1. We show that by searching over d_0 and d_1 we obtain a loss function which gives rise to a Bayes decision problem with a Bayes decision rule which has Type I error α and Type II error β at $\mu = \pm\delta$, and which minimizes our chosen objective function over the set of all decision rules. This Bayes decision rule can equally be viewed as an optimal stopping rule for our original frequentist problem.

In §4.6 we describe our improved method for objective function F_2: $E(N|\mu=\delta)$, while in §4.7 we indicate how to adapt our approach when minimizing other objective functions.

4.6 An Improved Method for the Computation of Optimal Two-Sided Group Sequential Tests.

Consider again the clinical trials problem outlined in §4.3 with a maximum of K equally sized groups of $2n$ patients. The random variable X_i $(i = 1, 2, \ldots, Kn)$ denotes the difference in response between the ith pair of patients and is normally distributed with unknown mean μ and known variance σ^2.

Here we wish to make a choice between the three decisions :

$$D_1^-: \mu < 0, \qquad D_0: \mu = 0 \qquad \text{and} \qquad D_1^+: \mu > 0.$$

Decisions D_1^- and D_1^+ can be made at any one of the K analyses, but we can only make decision D_0 at the Kth analysis.

We shall consider a family of Bayes decision theory problems with a common prior distribution for μ given by $\pi(-\delta) = \pi(0) = \pi(\delta) = 1/3$ with $\pi(\mu) = 0$ otherwise, and a common cost of sampling function given by $c(-\delta) = c(\delta) = 1$ with $c(\mu) = 0$ otherwise. Individual problems within the family differ in their loss functions, $L(D,\mu)$, which are indexed by a pair of parameters, d_0 (> 0) and d_1 (> 0). The general form of $L(D,\mu)$ is given by

$$L(D_0,-\delta) = L(D_1^+,-\delta) = d_1$$

$$L(D_1^-,0) = L(D_1^+,0) = d_0$$

$$L(D_1^-,\delta) = L(D_0,\delta) = d_1$$

with $L(D,\mu) = 0$ otherwise.

Suppose for the moment that d_0 and d_1 are fixed. Consider a general decision rule for the above problem which we shall denote by $\mathcal{B}$. Because our family of Bayes decision problems is symmetric about $\mu=0$ we shall only consider decision rules which are symmetric about zero. So $\mathcal{B}$ is of the general form :

At analysis i $(1 \leq i \leq K-1)$,

if $S_{in} \geq c_i$ stop entering patients on to the trial and make decision D_1^+;

if $S_{in} \leq -c_i$ stop entering patients on to the trial and make decision D_1^-;

if $|S_{in}| < c_i$ enter the next group of $2n$ patients on to the trial.

At analysis K,

if $S_{Kn} \geq c_K$ stop entering patients on to the trial and make decision D_1^+;

if $S_{Kn} \leq -c_K$ stop entering patients on to the trial and make decision D_1^-;

if $|S_{Kn}| < c_K$ stop entering patients on to the trial and make decision D_0.

The **risk** associated with $\mathcal{B}$, which we shall denote by $r(\mathcal{B},d_0,d_1)$, is defined as the sum of the total expected sampling cost of the trial **plus** the total expected loss through making a wrong decision. That is

$$\begin{aligned} r(\mathcal{B},d_0,d_1) \;=\; & c(-\delta)\,E(N|\mu=-\delta)\,\pi(-\delta) + c(\delta)\,E(N|\mu=\delta)\,\pi(\delta) \\ & + d_1\,\Pr(D_0 \cup D_1^+ | \mu=-\delta)\,\pi(-\delta) \\ & + d_0\,\Pr(D_1^- \cup D_1^+ | \mu=0)\,\pi(0) \\ & + d_1\,\Pr(D_1^- \cup D_0 | \mu=\delta)\,\pi(\delta) \end{aligned}$$

where $\Pr(D_i \cup D_j|\mu)$ denotes the probability of making either decision D_i or decision D_j under a treatment difference μ.

As $F_2 = E(N|\mu=\delta) = E(N|\mu=-\delta)$ for symmetric decision rules, it follows that

$$r(\mathcal{B},d_0,d_1) = \frac{1}{3}\{2F_2 + d_1 \Pr(D_0 \cup D_1^+|\mu=-\delta) + d_0 \Pr(D_1^- \cup D_1^+|\mu=0) + d_1 \Pr(D_1^- \cup D_0|\mu=\delta)\}.$$

The Bayes decision rule for our problem, $\mathcal{B}^*(d_0,d_1)$, minimizes this risk over the set of all decision rules, $\mathcal{S}$, i.e.

$$r(\mathcal{B}^*(d_0,d_1),d_0,d_1) = \min_{\mathcal{B}\in\mathcal{S}} \{r(\mathcal{B},d_0,d_1)\}.$$

We can compute $\mathcal{B}^*(d_0,d_1)$ by dynamic programming (the necessary computations are described in §4.8). Using numerical integration we can calculate the Type I and Type II errors of $\mathcal{B}^*(d_0,d_1)$. Suppose these errors are given by

$$\Pr(D_1^- \cup D_1^+|\mu=0) = \tilde{\alpha}$$

and

$$\Pr(D_0 \cup D_1^+|\mu=-\delta) = \Pr(D_1^- \cup D_0|\mu=\delta) = \tilde{\beta},$$

then clearly

$$r(\mathcal{B}^*(d_0,d_1),d_0,d_1) = \frac{1}{3}\{2F_2 + 2d_1\tilde{\beta} + d_0\tilde{\alpha}\}.$$

It follows from the definition of the Bayes rule that there can be no other decision rule with errors $\tilde{\alpha}$ and $\tilde{\beta}$ which attains a lower value of F_2.

Using the algorithm of Powell (1970), we search over (d_0,d_1) for a pir of loss parameters $(d_0^{(\alpha)},d_1^{(\beta)})$ giving rise to a Bayes decision theory problem with a Bayes rule, $\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)})$, which has errors α and β. Clearly

$$r(\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)}),d_0^{(\alpha)},d_1^{(\beta)}) = \frac{1}{3}\{2F_2 + 2d_1^{(\beta)}\beta + d_0^{(\alpha)}\alpha\}.$$

Again, from the definition of the Bayes rule, there can be no other decision rule for this problem with a smaller risk, i.e.

$$r(\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)}),d_0^{(\alpha)},d_1^{(\beta)}) = \min_{\mathcal{B}\in\mathcal{S}} \{r(\mathcal{B},d_0^{(\alpha)},d_1^{(\beta)})\}.$$

Moreover there can be no other decision rule with errors α and β which attains a lower value of F_2. By equating decisions D_1^-, D_0 and D_1^+ with the acceptance of hypotheses H_1^-, H_0 and H_1^+ respectively we have computed the optimal two-

sided group sequential test for our original frequentist problem.

4.7 Applying our Improved Method to Other Objective Functions.

Our improved method for the computation of optimal two-sided group sequential tests for F_1 is easily adapted for use with other objective functions. Here we provide details of the necessary adaptations for objective functions F_3: $E(N|\mu=2\delta)$ and F_5: $\int E(N|\mu)\,\delta^{-1}\,\varphi(\mu/\delta)\,d\mu$.

Consider the minimization of F_3. A suitable family of Bayes decision theory problems has the common prior distribution for μ given by $\pi(-2\delta) = \pi(-\delta) = \pi(0) = \pi(\delta) = \pi(2\delta) = 1/5$ with $\pi(\mu) = 0$ otherwise, and a common cost of sampling function given by $c(-2\delta) = c(2\delta) = 1$ with $c(\mu) = 0$ otherwise. The general form of the loss function, $L(D,\mu)$, is identical to that given in §4.6, with, in particular, $L(D,2\delta) = L(D,-2\delta) = 0$ for any decision, D.

For fixed loss parameters, d_0 and d_1, the **risk** of a given decision rule, $\mathcal{B}$, is denoted by $r(\mathcal{B},d_0,d_1)$ and equals

$$\begin{aligned} & c(-2\delta)\,E(N|\mu=-2\delta)\,\pi(-2\delta) + c(2\delta)\,E(N|\mu=2\delta)\,\pi(2\delta) \\ & \quad + d_1\,\Pr(D_0\cup D_1{}^+|\mu=-\delta)\,\pi(-\delta) \\ & \quad + d_0\,\Pr(D_1{}^-\cup D_1{}^+|\mu=0)\,\pi(0) \\ & \quad + d_1\,\Pr(D_1{}^-\cup D_0|\mu=\delta)\,\pi(\delta) \end{aligned}$$

That is,

$$\begin{aligned} r(\mathcal{B},d_0,d_1) = \frac{1}{5}\{ & 2F_3 + d_1\,\Pr(D_0\cup D_1{}^+|\mu=-\delta) \\ & + d_0\,\Pr(D_1{}^-\cup D_1{}^+|\mu=0) + d_1\,\Pr(D_1{}^-\cup D_0|\mu=\delta)\}. \end{aligned}$$

The Bayes decision rule for this problem, $\mathcal{B}^*(d_0,d_1)$, minimizes this risk over the set of all decision rules, $\mathcal{S}$, i.e.

$$r(\mathcal{B}^*(d_0,d_1),d_0,d_1) = \min_{\mathcal{B}\in\mathcal{S}}\{r(\mathcal{B},d_0,d_1)\}.$$

Suppose $\mathcal{B}^*(d_0,d_1)$ has errors given by

$$\Pr(D_1{}^-\cup D_1{}^+|\mu=0) = \tilde{\alpha}$$

and

$$\Pr(D_0\cup D_1{}^+|\mu=-\delta) = \Pr(D_1{}^-\cup D_0|\mu=\delta) = \tilde{\beta},$$

then clearly

$$r(\mathcal{B}^*(d_0,d_1),d_0,d_1) = \frac{1}{5}\{2F_3 + 2d_1\,\tilde{\beta} + d_0\,\tilde{\alpha}\},$$

and, from the definition of the Bayes rule, there can be no other decision rule with errors $\tilde{\alpha}$ and $\tilde{\beta}$ which attains a lower value of F_3.

By searching over (d_0,d_1) we obtain a pair of loss parameters, $d_0^{(\alpha)}$ and $d_1^{(\beta)}$, giving a Bayes decision theory problem with an associated Bayes rule, $\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)})$, which has errors α and β. Clearly,

$$r(\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)})) = \frac{1}{5}\{2F_3 + 2d_1\,\beta + d_0\,\alpha\}.$$

By the definition of the Bayes decision rule there can be no other decision rule for this problem with a smaller risk, i.e.

$$r(\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)}),d_0^{(\alpha)},d_1^{(\beta)}) = \min_{\mathcal{B}\in\mathcal{S}}\{r(\mathcal{B},d_0^{(\alpha)},d_1^{(\beta)})\}.$$

Moreover there can be no other decision rule with errors α and β which attains a lower value of F_3. By equating decisions D_1^-, D_0 and D_1^+ with the acceptance of the hypotheses H_1^-, H_0 and H_1^+ we obtain the optimal two-sided group sequential test for our original frequentist problem.

For the minimization of F_5: $\int E(N|\mu)\,\delta^{-1}\,\varphi(\mu/\delta)\,d\mu$ we consider the family of Bayes decision theory problems with the common prior distribution:

$$\pi(\mu) = \begin{cases} 1/4 & \text{if } \mu = -\delta, 0 \text{ or } \delta \\ (1/4)\delta^{-1}\varphi(\mu/\delta) & \text{otherwise} \end{cases}$$

and the common cost of sampling function

$$c(\mu) = \begin{cases} 1 \text{ if } \mu \neq -\delta, 0 \text{ or } +\delta \\ 0 \text{ otherwise.} \end{cases}$$

The general form of the loss function is identical to that given in §4.6, with, in particular, $L(D,\mu) = 0$ for $\mu \neq 0, \pm\delta$.

For fixed d_0 and d_1, the **risk** of a general decision rule, $\mathcal{B}$, is denoted by $r(\mathcal{B},d_0,d_1)$, and is equal to

$$\int c(\mu)\,E(N|\mu)\,\pi(\mu)\,d\mu + d_1\,\Pr(D_0 \cup D_1^+|\mu=-\delta)\,\pi(-\delta)$$

$$+\, d_0\,\Pr(D_1^- \cup D_1^+|\mu=0)\,\pi(0) + d_1\,\Pr(D_1^- \cup D_0|\mu=\delta)\,\pi(\delta).$$

That is,

$$r(\mathcal{B},d_0,d_1) = \frac{1}{4}\Big\{\int E(N|\mu)\,\delta^{-1}\,\varphi(\mu/\delta)\,d\mu +$$

$$d_1\,\Pr(D_0\cup D_1^{+}|\mu=-\delta) + d_0\,\Pr(D_1^{-}\cup D_1^{+}|\mu=0) + d_1\,\Pr(D_1^{-}\cup D_0|\mu=\delta)\Big\}$$

$$= \frac{1}{4}\{F_5+d_1\Pr(D_0\cup D_1^{+}|\mu=-\delta)+d_0\Pr(D_1^{-}\cup D_1^{+}|\mu=0)+d_1\Pr(D_1^{-}\cup D_0|\mu=\delta)\}.$$

The rest of the logic is analogous to that for objective functions F_2 and F_3 described earlier. In particular, we search over (d_0,d_1) for a pair of loss parameters $d_0^{(\alpha)}$ and $d_1^{(\beta)}$, giving rise to a Bayes decision theory problem with an associated Bayes rule, $\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)})$, which has errors α and β. By the definition of the Bayes rule there can be no other decision rule for this problem with a smaller risk, i.e.

$$r(\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)}),d_0^{(\alpha)},d_1^{(\beta)}) = \min_{\mathcal{B}\in S}\{r(\mathcal{B},d_0^{(\alpha)},d_1^{(\beta)})\}.$$

Moreover there can be no other decision rule with errors α and β which attains a lower value of F_5. By equating decisions D_1^{-}, D_0 and D_1^{+} with the acceptance of the hypotheses H_1^{-}, H_0 and H_1^{+} we obtain the optimal two-sided group sequential test for our original frequentist problem.

Clearly our improved method for computing optimal two-sided group sequential tests is easily extended to other objective functions.

4.8 The Dynamic Programming Algorithm.

As has already been mentioned, for any given Bayes decision theory problem the Bayes rule is computed using dynamic programming. In this section we describe the dynamic programming algorithm for a general problem. Throughout we shall assume that the Bayes rule is monotone. That is at stage i $(i=1,2,\ldots,K-1)$ it is optimal to make decision D_1^{+} if $S_{in}\geq c_i$ and decision D_1^{-} if $S_{in}\leq -c_i$, while for $-c_i<S_{in}<c_i$ it is optimal to sample the next group of n observations. Further, at stage K it is optimal to make decision D_1^{+} if $S_{Kn}\geq c_K$, decision D_0 if $-c_K<S_{Kn}<c_K$ and decision D_1^{-} if $S_{Kn}\leq -c_K$. Numerical checks

support our assumption that $\mathcal{B}^*(d_0,d_1)$ is monotone.

Consider again the problem described in §4.6 with a maximum of K groups of n observations available for choosing between the decisions

$$D_1^-: \mu < 0, \quad D_0: \mu = 0 \quad \text{and} \quad D_1^+: \mu > 0.$$

This time we have a general prior distribution, $\pi(\mu)$, defined over some parameter space M, and a general cost of sampling function, $c(\mu)$, also defined over M. Finally we have a loss function, $L(D,\mu)$, which is of the same general form as that given in §4.6.

Suppose that the loss parameters d_0 and d_1 are fixed. Letting $p^{(i)}(\mu|x)$ denote the current posterior distribution of μ at analysis i $(1 \le i \le K)$, given that $S_{in} = x$, the loss from making decision D_1^+ equals

$$d_1\, p^{(i)}(-\delta|x) + d_0\, p^{(i)}(0|x)$$

and the loss from making decision D_1^- equals

$$d_0\, p^{(i)}(0|x) + d_1\, p^{(i)}(\delta|x).$$

At analysis K the loss from making decision D_0 equals

$$d_1\, \{p^{(K)}(-\delta|x) + p^{(K)}(\delta|x)\}.$$

If we let $\gamma^{(i)}(x)$ denote the minimum loss from stopping at stage i $(i = 1, 2, \ldots, K)$ and making a definite decision, it follows that

$$\gamma^{(K)}(x) = \min\big[\{d_0\, p^{(K)}(0|x) + d_1\, p^{(K)}(\delta|x)\}, d_1\, \{p^{(K)}(-\delta|x) + p^{(K)}(\delta|x)\},$$
$$\{d_1\, p^{(K)}(-\delta|x) + d_0\, p^{(K)}(0|x)\}\big]$$

while, for $i \le K-1$,

$$\gamma^{(i)}(x) = \min\big[\{d_0\, p^{(i)}(0|x) + d_1\, p^{(i)}(\delta|x)\}, \{d_1\, p^{(i)}(-\delta|x) + d_0\, p^{(i)}(0|x)\}\big].$$

Further, let $\beta^{(i)}(x)$ denote the minimum additional risk from sampling the $(i+1)$st group $(i = 1, 2, \ldots, K-1)$ and then proceeding optimally. Denoting the c.d.f. of $S_{(i+1)n}$ given that $S_{in} = x$ by $F^{(i+1)}(S_{(i+1)n}|x)$, we have

$$\beta^{(K-1)}(x) = n \sum_{\mu \in M} c(\mu) p^{(K-1)}(\mu|x) + \int_{S_{Kn}} \gamma^{(K)}(S_{Kn})\, dF^{(K)}(S_{Kn}|x)$$

and, for $i < K-1$,

$$\beta^{(i)}(x) = n \sum_{\mu \in M} c(\mu) p^{(i)}(\mu|x)$$

$$+ \int_{S_{(i+1)n}} \min \{\beta^{(i+1)}(S_{(i+1)n}), \gamma^{(i+1)}(S_{(i+1)n})\} \, dF^{(i+1)}(S_{(i+1)n}|x)$$

where the summation signs in the above equations are replaced by a mixture of sums and integrals for objective functions such as F_5.

We compute the critical values of $\mathcal{B}^*(d_0, d_1)$ by starting at the Kth analysis and working back.

At analysis K it is optimal to make decision D_1^+ for all x (≥ 0) such that

$$d_1 \, p^{(K)}(-\delta|x) + d_0 \, p^{(K)}(0|x) < d_1 \{p^{(K)}(-\delta|x) + p^{(K)}(\delta|x)\} \quad (4.8.1)$$

and to make decision D_0 for all x (≥ 0) such that

$$d_1 \{p^{(K)}(-\delta|x) + p^{(K)}(\delta|x)\} < d_1 \, p^{(K)}(-\delta|x) + d_0 \, p^{(K)}(0|x). \quad (4.8.2)$$

Using the bisection method we obtain the Kth critical value, c_K, as the solution, for $x \geq 0$, of the equation

$$d_1 \{p^{(K)}(-\delta|x) + p^{(K)}(\delta|x)\} = d_1 \, p^{(K)}(-\delta|x) + d_0 \, p^{(K)}(0|x). \quad (4.8.3)$$

Clearly it is possible that equation (4.8.3) does not possess a solution for $x \geq 0$. To avoid our algorithm getting in to computational difficulties we can build a simple check in to our program. The check tests whether inequality (4.8.2) holds for $x = 0$. If this is not the case we set $c_K = 0$ and go back to the $(K-1)$st analysis. Similar checks can be built in at other analyses.

The symmetry inherent in our problem makes it optimal to choose decision D_1^- if $S_{Kn} \leq -c_K$, and decision D_0 if $-c_K < S_{Kn} \leq 0$.

At analysis $K-1$ it is optimal to make decision D_1^+ for all $S_{(K-1)n} = x$ (≥ 0) such that

$$d_1 \, p^{(K-1)}(-\delta|x) + d_0 \, p^{(K-1)}(0|x) < \beta^{(K-1)}(x)$$

and to decide to sample the Kth group of observations and then to proceed optimally for all x (≥ 0) such that

$$\beta^{(K-1)}(x) < d_1 \, p^{(K-1)}(-\delta|x) + p^{(K-1)}(0|x).$$

Here $\beta^{(K-1)}(x)$ is given by

$$n \sum_{\mu \in M} c(\mu) p^{(K-1)}(\mu|x)$$

$$+ d_1 \Pr_{-\delta}(S_{Kn} \geq c_K \mid x)\, p^{(K-1)}(-\delta|x) + d_0 \Pr_0(S_{Kn} \geq c_K \mid x)\, p^{(K-1)}(0|x)$$

$$+ d_1 \{\Pr_{-\delta}(|S_{Kn}| \leq c_K \mid x)\, p^{(K-1)}(-\delta|x) + \Pr_{\delta}(|S_{Kn}| \leq c_K \mid x)\, p^{(K-1)}(\delta|x)\}$$

$$+ d_0 \Pr_0(S_{Kn} \leq -c_K \mid x)\, p^{(K-1)}(0|x) + d_1 \Pr_{\delta}(S_{Kn} \leq -c_K \mid x)\, p^{(K-1)}(\delta|x).$$

Again we use the bisection method to obtain the $(K-1)$st critical value, c_{K-1} as the solution, for $x \geq 0$, of equation (4.8.4)

$$d_1\, p^{(K-1)}(-\delta|x) + d_0\, p^{(K-1)}(0|x) = \beta^{(K-1)}(x). \tag{4.8.4}$$

The symmetry inherent in our problem makes it optimal to choose D_1^- if $S_{(K-1)n} \leq -c_{K-1}$ and to continue sampling if $-c_{K-1} < S_{(K-1)n} \leq 0$.

At analysis $K-2$ it is optimal to make decision D_1^+ for all $S_{(K-2)n} = x$ (≥ 0) such that

$$d_1\, p^{(K-2)}(-\delta|x) + d_0\, p^{(K-2)}(0|x) < \beta^{(K-2)}(x)$$

and to decide to enter the $(K-1)$st group of patients on to the trial and then to proceed optimally for all x (≥ 0) such that

$$\beta^{(K-2)}(x) < d_1\, p^{(K-2)}(-\delta|x) + d_0\, p^{(K-2)}(0|x).$$

Here $\beta^{(K-2)}(x)$ is given by

$$n \sum_{\mu \in M} c(\mu) p^{(K-2)}(\mu|x)$$

$$+ d_1 \Pr_{-\delta}(S_{(K-1)n} \geq c_{K-1} \mid x)\, p^{(K-2)}(-\delta|x) + d_0 \Pr_0(S_{(K-1)n} \geq c_{K-1} \mid x)\, p^{(K-2)}(0|x)$$

$$+ d_0 \Pr_0(S_{(K-1)n} \leq -c_{K-1} \mid x)\, p^{(K-2)}(0|x) + d_1 \Pr_{\delta}(S_{(K-1)n} \leq -c_{K-1} \mid x)\, p^{(K-2)}(\delta|x)$$

$$+ \int_{-c_{K-1}}^{c_{K-1}} \beta^{(K-1)}(S_{(K-1)n})\, dF^{(K-1)}(S_{(K-1)n}|x)$$

Again the bisection method gives the $(K-2)$nd critical value, c_{K-2}, as the solution, for $x \geq 0$, of equation (4.8.5)

$$d_1\, p^{(K-2)}(-\delta|x) + d_0\, p^{(K-2)}(0|x) = \beta^{(K-2)}(x). \tag{4.8.5}$$

The symmetrical nature of our problem makes it optimal to choose D_1^- if

$S_{n(K-2)} \leq -c_{K-2}$ and to continue sampling if $-c_{K-2} < S_{n(K-2)} \leq 0$.

We work back to the first analysis in a similar fashion. The error probabilities of the resulting decision rule are then computed and a test for convergence conducted. If convergence has not been achieved the error probabilities are fed into Powell's method and a new pair (d_0 , d_1) obtained.

4.9 Results and Discussion.

Tables 4.2-4.7 give the minima of $F_2: E(N|\mu=\delta)$, $F_3: E(N|\mu=2\delta)$ and $F_5: \int E(N|\mu)\delta^{-1}\varphi(\mu/\delta)\, d\mu$ expressed as percentages of the fixed sample size, N_f, for $\alpha=0.05$, $\beta=0.05$ and 0.1, K = 2, 3, 4, 5 and 10 and t, the ratio of the maximum sample size of the sequential test to N_f, equal to 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6. The minima are independent of σ^2 and δ (for verification of this fact see Appendix 5.2).

In each table the entry for $K=10$ and $t=1.6$ is missing. So, although it is possible to compute feasible tests for each of these problems, there does not exist a Bayes decision theory problem giving rise to a Bayes rule with the required errors. The reason for this is almost certainly linked to the fact that decision D_0 is only open to the experimenter at the Kth analysis. For the inner wedge problems considered in §5, where any of the decisions D_1^-, D_0 and D_1^+ can be made at each analysis, optimal feasible tests were always obtained.

Tables 4.2 and 4.3 refer to the minimization of $F_2: E(N|\mu=\delta)$ for $\alpha=0.05$ and $\beta=0.05$ and 0.1 respectively.

Table 4.2. Minima of $F_2: E(N|\mu=\delta)$ expressed as percentages of the fixed sample size, N_f, for $\alpha=0.05$, $\beta=0.05$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6, and K = 2, 3, 4, 5 and 10.

					t				
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	78.6	72.6	**71.8**	72.4	73.6	76.8	80.4	84.3	88.3
3	74.0	67.2	65.2	**64.8**	65.0	66.3	68.1	70.1	72.3
4	71.6	64.8	62.6	61.9	**61.8**	62.5	63.7	65.1	66.6
5	70.1	63.3	61.0	60.3	**60.1**	60.6	61.6	62.7	63.8
10	67.1	60.3	57.9	57.1	**56.8**	57.0	57.6	58.4	

In considering Table 4.2 we begin by looking at group sequential tests with $K=2$. Each of these tests leads to large savings in F_2 compared with the fixed sample size test. The largest saving occurs with $t=1.1$ and equals 28.2% of the fixed sample size. Even if logistical considerations limit us to a test with $K=2$ and $t=1.01$ the expected saving is 21.4% of the fixed sample size.

For fixed t gains in efficiency increase with K, however the rate of gain in efficiency decreases with K. For example with $t=1.2$ the expected gain in efficiency in doubling the maximum number of groups from 5 to 10 is only 3.3% of the fixed sample size.

For fixed K, the minimum of F_2 over tabulated values of t is shown in bold type. We see that even with $K=10$ there would appear to be no point in designing experiments with t much greater than 1.2. Note that, as with the one-sided optimal tests of §3.9, the minima of F_2 are a U-shaped function of t for K fixed.

Many of the comments made concerning Table 4.2 apply equally to Table 4.3. Again the largest gains in efficiency occur when going from a fixed sample size test to a group sequential test with a maximum of 2 groups. For instance with $t=1.1$ and $K=2$ the expected saving over the fixed sample size test is

22.4% of N_f. Even with $t = 1.01$ and $K = 2$ this expected saving is 16.1% of N_f. Most of the gains in efficiency occur with $K \leq 5$. As can be seen doubling K from 5 to 10 results in only moderate gains in efficiency.

As with Table 4.2 the minimum of F_2 over tabulated values of t for K fixed is shown in bold type. Here, there would appear to be no point in considering tests with $t > 1.15$.

Table 4.3. Minima of $F_2 : E(N|\mu=\delta)$ expressed as percentages of the fixed sample size, N_f, for $\alpha=0.05$, $\beta=0.1$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6, and K = 2, 3, 4, 5 and 10.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	83.9	78.4	**77.6**	78.2	79.4	82.6	86.2	90.0	94.0
3	79.9	73.9	72.3	**72.1**	72.5	74.2	76.3	78.6	80.9
4	77.7	71.7	70.0	**69.6**	69.8	71.0	72.7	74.4	76.2
5	76.3	70.4	68.6	**68.1**	68.3	69.3	70.8	72.3	73.9
10	73.6	67.6	65.7	**65.1**	65.2	65.9	67.1	68.4	

Tables 4.4 and 4.5 give the minima of $F_3 : E(N|\mu=2\delta)$ for $\alpha = 0.05$ and $\beta = 0.05$ and 0.1 respectively. Obviously it is when treatment differences are large that the ethical argument for stopping a trial early is strongest. Both tables illustrate the advantage of an optimal group sequential test when treatment differences are large.

Table 4.4. Minima of $F_3: E(N|\mu=2\delta)$ expressed as percentages of the fixed sample size, N_f, for $\alpha=0.05$, $\beta=0.05$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6, and K = 2, 3, 4, 5 and 10.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	**50.9**	52.6	55.0	57.5	60.0	65.0	70.0	75.0	80.0
3	36.8	**36.2**	37.3	38.7	40.3	43.5	46.7	50.0	53.4
4	32.0	**29.4**	29.5	30.2	31.1	33.2	35.4	37.8	40.2
5	30.0	26.3	**25.6**	25.8	26.2	27.5	29.1	30.8	32.6
10	26.7	22.1	20.5	19.7	19.4	**19.2**	19.4	19.8	

As can be seen gains in efficiency are extremely impressive. For example with $K=2$ it is almost certain that the trial will stop at the first analysis under $\mu=\pm 2\delta$. While with $K=10$ the expected sample size is only about 20% of the fixed sample size for each value of β and K.

In both tables the largest gains in efficiency occur when going from a fixed sample size test to a group sequential test with $K=2$ There is a stronger case for considering tests with a "large" maximum number of groups here than was the case with optimal tests for objective function F_2. For fixed t there are quite substantial gains to be made in going from a 5 to a 10 group design. This is because, under $\mu=\pm 2\delta$, there is a high probability that the test will stop at the first analysis and, for fixed t, the first analysis occurs after twice as many observations when $K=5$ compared with when $K=10$.

Table 4.5. Minima of $F_3: E(N|\mu=2\delta)$ expressed as percentages of the fixed sample size, N_f, for α=0.05, β=0.1, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6, and number of groups, K = 2, 3, 4, 5 and 10.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	**52.0**	53.0	55.2	57.6	60.1	65.0	70.0	75.0	80.0
3	40.3	**38.0**	38.5	39.6	41.0	43.9	47.0	50.2	53.5
4	36.7	32.5	**31.8**	32.1	32.7	34.4	36.3	38.5	40.7
5	35.0	30.0	28.7	**28.4**	28.6	29.4	30.7	32.1	33.7
10	31.7	26.5	24.5	23.5	23.0	22.6	**22.6**	22.8	

Tables 4.6 and 4.7 refer to the minimization of objective function $F_5: \int E(N|\mu)\delta^{-1}\varphi(\mu/\delta)\,d\mu$ for $\alpha = 0.05$ and $\beta = 0.05$ and 0.1 respectively.

Table 4.6. Minima of $F_5: \int E(N|\mu)\delta^{-1}\varphi(\mu/\delta)\,d\mu$ expressed as percentages of the fixed sample size, N_f, for α=0.05, β=0.05, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6, and K = 2, 3, 4, 5 and 10.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	84.5	**83.6**	85.3	84.5	90.5	96.4	102.6	108.8	115.1
3	81.1	**79.2**	80.1	81.9	84.1	89.0	94.1	99.4	104.7
4	79.6	**77.4**	78.0	79.6	81.6	86.0	90.7	95.6	100.4
5	78.7	**76.4**	77.0	78.4	80.3	84.5	89.0	93.7	98.3
10	76.9	**74.6**	76.9	76.4	78.1	82.0	86.3	90.6	

Gains in efficiency here are less impressive than was the case with objective functions F_2 and F_3. Indeed for certain tests in each table the expected sample size exceeds the fixed sample size for this problem. The reason for this is linked with the prior distribution for μ, which places a large proportion of its total density at or around $\mu = 0$, where the probability of stopping early is small.

The most impressive gains in efficiency occur when t is small. For example with $K = 2$, $\beta = 0.05$ and $t = 1.05$ the expected saving over the fixed sample size test is 16.4% of N_f. With $K = 5$, $\beta = 0.05$ and $t = 1.05$ this saving has increased to 23.6% of N_f. Expected gains in efficiency from doubling K from 5 to 10 are small here. For instance with $\beta = 0.1$ and $t = 1.05$ the expected gain in efficiency in going from a design with $K = 5$ to one with $K = 10$ is just 1.7% of the fixed sample size.

In practice there would appear to be little point in employing an optimal test for F_5 with $t > 1.05$ or $K > 5$.

Table 4.7. Minima of $F_5 : \int E(N|\mu)\delta^{-1}\varphi(\mu/\delta)\,d\mu$ expressed as percentages of the fixed sample size, N_f, for α=0.05, β=0.1, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6, and number of groups, K = 2, 3, 4, 5 and 10.

					t				
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	86.9	**86.1**	87.8	90.3	93.1	99.2	105.4	111.8	118.2
3	84.0	**82.3**	83.4	85.3	87.6	92.6	98.0	103.5	108.9
4	82.7	**80.8**	81.6	83.4	85.4	90.1	95.1	100.2	105.3
5	81.9	**79.9**	80.7	82.3	84.3	88.8	93.6	98.5	103.4
10	80.2	**78.2**	79.0	80.5	82.4	86.6	91.2	95.8	

We have not considered the minimization of $F_1 = E(N|\mu{=}0)$ so far in this chapter. Under $\mu = 0$ the probability of stopping early is at most α (from equation (4.2.1)). Hence a lower bound on F_1 is $n\alpha + nK(1-\alpha)$, which, even for fairly

small maximum sample sizes, will exceed the fixed sample size. For example with $\alpha=0.05$, $\beta=0.1$, $t=1.01$ and K = 2, 3, 4, 5 and 10 the minima of F_1 expressed as percentages of the relevant fixed sample size are given in Table 4.8. The results are independent of δ and σ^2.

Table 4.8. Minima of F_1 expressed as percentages of the fixed sample size, N_f, for tests with $\alpha=0.05$, $\beta=0.1$, $t=1.01$ and K = 2, 3, 4, 5 and 10. Results given are independent of δ and σ^2.

K	2	3	4	5	10
$100F_1/N_f$	100.67	100.61	100.57	100.55	100.48

For each of the 5 tests considered in Table 4.8 the expected sample size under $\mu=0$ is greater than the fixed sample size. Indeed, at best, $E(N|\mu=0)$ is only 0.5% of N_f less than the maximum sample size. If minimizing the expected number of patients entering a trial under $\mu=0$ is important we would recommend the use of one of the inner wedge tests described in §5.

The results for designs with more than 10 groups, for other objective functions and for other values of α and β are not given here. It suffices to say that our method for the computation of optimal two-sided group sequential tests is easily extended to deal with such problems.

4.10 Examples.

In this section we give three examples of optimal two-sided group sequential tests. All of the examples are based on the same hypothesis testing problem which we now describe.

Suppose $X_1, X_2, \ldots, X_{Kn}$ are independent normal random variables with unknown mean μ and unit variance, and that we wish to test

$$H_1^-: \mu<0 \quad vs \quad H_0: \mu=0 \quad vs \quad H_1^+: \mu>0$$

with error rates

$$\Pr(\mathcal{A}_1^- \cup \mathcal{A}_1^+ | \mu = 0) = 0.05 \tag{4.10.1}$$

$$\Pr(\mathcal{A}_0 \cup \mathcal{A}_1^+ | \mu = -0.5) = 0.1 \tag{4.10.2}$$

and

$$\Pr(\mathcal{A}_1^- \cup \mathcal{A}_0 | \mu = 0.5) = 0.1. \tag{4.10.3}$$

The fixed sample size test for this problem would require 42.03 patients on each treatment arm. In practice, of course, we would assign 42 patients to each treatment with the result that the power of detecting the treatment differences $\mu = \pm 0.5$ would be slightly less than 0.9.

For our first example we consider the group sequential test with a maximum of 5 groups of 20 patients which minimizes F_2 subject to the error constraints (4.10.1)-(4.10.3). The standardized critical values for this test, c_i' $(=c_i/\sqrt{in})$, are given by: $c_1' = 2.537$, $c_2' = 2.350$, $c_3' = 2.369$, $c_4' = 2.426$ and $c_5' = 2.381$. The corresponding nominal significance levels are: $\alpha_1 = 0.011$, $\alpha_2 = 0.019$, $\alpha_3 = 0.018$, $\alpha_4 = 0.015$ and $\alpha_5 = 0.017$. As can be seen the α_i's are not monotonically increasing as one might expect. This point was also taken up by Pocock (1982) who noted the 'curious' tendency for α_4 to be smaller than α_3 and/or α_2 in optimal 5 group designs. We note however that there is no theoretical reason for expecting increasing nominal levels.

For our first example we have $E(N|\mu = 0.5) = 28.7$ (remember this represents the expected number of patients on each treatment) which is 68.2% of the corresponding fixed sample size. Clearly this is a substantial saving in terms of both financial and human resources. The maximum of the expected sample size function (at $\mu = 0$) is very close to 50. Indeed $|\mu|$ must exceed 0.2 before we can expect our optimal test to require fewer than the 42 patients needed by the fixed sample size test. (If minimizing expected sample size for $|\mu|$ small is a priority we would strongly recommend the use of one of the inner wedge designs of §5).

As a second example of an optimal group sequential test we consider the minimization of $F_3 : E(N|\mu = 2\delta)$ for the same problem as in our first example. The standardized critical values for this test are: $c_1' = 2.203$, $c_2' = 2.626$, $c_3' = 2.949$, $c_4' = 3.095$ and $c_5' = 2.309$. The corresponding nominal significance

levels are given by: $\alpha_1 = 0.028$, $\alpha_2 = 0.009$, $\alpha_3 = 0.003$, $\alpha_4 = 0.002$ and $\alpha_5 = 0.021$. It is interesting to note that more than half of the total Type I error is spent at the first analysis while almost all the remaining Type I error is spent at the final analysis. This is in marked contrast to the first example.

Differences between $E(N|\mu)$ here compared with our first example are small, although the optimal test for F_2 is better for $|\mu|$ close to 0. Under $\mu = \pm 1$ the second test is optimal, of course, with $E(N|\mu = 1)$ = 12.0 or 28.5% of the corresponding fixed sample size.

For our third example we consider the group sequential test which minimizes F_5 for the same problem as in our earlier two examples. The standardized critical values for this test are given by: $c_1' = 2.463$, $c_2' = 2.386$, $c_3' = 2.403$, $c_4' = 2.440$ and $c_5' = 2.374$. The corresponding nominal significance levels are: $\alpha_1 = 0.014$, $\alpha_2 = 0.017$, $\alpha_3 = 0.016$, $\alpha_4 = 0.015$ and $\alpha_5 = 0.018$.

The operating characteristic function and $\Pr(\mathcal{A}_1^+|\mu)$ are almost identical for all three examples. This demonstrates that the frequentist, while only specifying errors at three points of the parameter space, effectively defines error functions over the entire parameter space.

4.11 A Comparison of Two-Sided Group Sequential Tests.

In this section we compare a selection of our optimal two-sided group sequential tests with 5 of the tests proposed in the literature and described in §§4.4 and 4.5. Comparisons are made in terms of the expected sample size under $\mu = \delta$ and the maximum sample size for each test.

For designs with K = 2, 3, 4, 5 and 10, $\alpha = 0.05$ and $\beta = 0.05$, Table 4.9 gives $F_2: E(N|\mu=\delta)$ and the maximum sample size, nK, both expressed as percentages of the relevant fixed sample size, N_f, for the tests of O'Brien and Fleming (OBF); Wang & Tsiatis (WT) with $\Delta = 0.25$; Pocock; Fleming, Harrington & O'Brien (FHOB) with $\gamma = 0.25$; Haybittle/Peto and the optimal test (F_2^*), which minimizes $E(N|\mu=\delta)$ over both feasible stopping rules **and** group sizes. The results of Table 4.9 are independent of both δ and σ^2.

In terms of minimizing F_2, F_2^* is, of course, optimal for each value of K. However Pocock's test is very close to optimal, never being more than 0.8% of

the fixed sample size above the overall minimum. The other 4 tests, in decreasing order of efficiency, are: WT, FHOB, OBF and Haybittle/Peto. Even the Haybittle/Peto tests, however, with their extremely simple stopping rule, show reasonable gains in efficiency over the fixed sample size test.

Table 4.9. $E(N|\mu=\delta)$ expressed as a percentage of the fixed sample size, N_f, and, in parentheses, t, the maximum sample size expressed as a proportion of N_f for the optimal test, F_2^*, OBF, WT ($\Delta=0.25$), Pocock, FHOB ($\gamma=0.25$) and the Haybittle/Peto test with $\alpha=0.05$, $\beta=0.05$ and K = 2, 3, 4, 5 and 10.

K	F_2^*	OBF	WT $\Delta=0.25$	Pocock	FHOB $\gamma=0.25$	Haybittle / Peto
2	71.8	80.2	74.0	71.8	75.1	83.9
	(1.10)	(1.01)	(1.03)	(1.09)	(1.02)	(1.00)
3	64.8	74.8	68.2	64.9	71.3	77.8
	(1.15)	(1.01)	(1.04)	(1.14)	(1.02)	(1.01)
4	61.8	71.3	65.5	61.9	69.8	74.4
	(1.18)	(1.02)	(1.05)	(1.17)	(1.02)	(1.01)
5	60.1	69.4	63.8	60.2	68.9	72.2
	(1.20)	(1.02)	(1.06)	(1.19)	(1.02)	(1.01)
10	56.7	65.7	60.1	57.5	67.4	67.1
	(1.22)	(1.03)	(1.07)	(1.25)	(1.02)	(1.03)

In terms of having a low maximum sample size the Haybittle/Peto tests can be seen to be the best. Even with 10 analyses it only requires at most 3% more observations than the single sample test. The OBF test is almost as good with $t \leq 1.03$ for $K \leq 10$. The optimal test and the Pocock test require the largest

maximum sample sizes. However it should be noted that optimal tests could be designed with the same values of t as Haybittle/Peto and with substantial savings in $E(N|\mu=\delta)$.

Table 4.10 is identical to Table 4.9 except that $\beta=0.1$. Most of the comments made concerning Table 4.9 are equally applicable here. It is interesting to note that the optimality of the Pocock test has declined slightly at $K=5$ and 10 with the increase in the Type II error. Conversely the OBF, Wang & Tsiatis, FHOB and Haybittle tests are closer to being optimal.

Table 4.10. $E(N|\mu=\delta)$ expressed as a percentage of the fixed sample size, N_f, and, in parentheses, t the maximum sample size expressed as a proportion of N_f, for the optimal test, $F_2{}^*$, OBF, WT ($\Delta=0.25$), Pocock, FHOB ($\gamma=0.25$) and the Haybittle/Peto test with $\alpha=0.05$, $\beta=0.1$ and K = 2, 3, 4, 5 and 10.

K	$F_2{}^*$	OBF	WT $\Delta=0.25$	Pocock	FHOB $\gamma=0.25$	Haybittle / Peto
2	77.6 (1.10)	85.0 (1.01)	78.9 (1.03)	77.6 (1.10)	80.5 (1.03)	88.2 (1.00)
3	72.1 (1.14)	79.5 (1.01)	74.1 (1.04)	72.1 (1.15)	77.5 (1.02)	83.4 (1.01)
4	69.6 (1.15)	76.4 (1.02)	71.6 (1.06)	69.7 (1.18)	76.3 (1.02)	80.6 (1.01)
5	68.1 (1.16)	74.7 (1.02)	70.1 (1.06)	68.5 (1.21)	75.6 (1.02)	78.8 (1.01)
10	65.1 (1.17)	71.3 (1.03)	66.8 (1.08)	66.6 (1.27)	74.4 (1.02)	74.5 (1.03)

4.12 The Repeated Confidence Interval Approach.

In this section we describe how the optimal group sequential tests of §§ 3 and 4 may be combined with the repeated confidence interval method of Jennison & Turnbull (1984), (1989) to give a flexible and optimal procedure for use in clinical trials.

Throughout §§3 and 4 it has been assumed that a study will only be stopped when a pre-defined stopping boundary is crossed. Of course in practice the decision to stop a trial is far more complex than this as it will often depend on such issues as the side-effects of the treatments and the quality of life for patients on the experiment. To quote Jennison & Turnbull (1984) '...As a result there is no guarantee that a particular stopping rule will be adhered to, nor is this desirable'.

Unfortunately frequentist properties, such as the probability of making a correct decision, depend on a strict adherence to a pre-defined stopping rule. As an alternative, more flexible, frequentist analysis of interim results Jennison & Turnbull (1984), (1989) have suggested constructing repeated confidence intervals (RCIs) for the parameter of interest. A suitable choice of these RCIs ensures that the probability that they all contain the true parameter value is acceptably high. At any analysis a decision to stop the study can be made on the basis of the current interval. However, if the study is stopped for any other reason, such as those cited earlier, the RCIs remain valid.

As an example of the RCI approach consider the clinical trial outlined in §4.3 with two treatments, A and B, and a maximum of K groups of $2n$ patients available for entry on to the study. Again let X_i denote the difference in response between the ith patients on treatments A and B respectively. We assume that $X_i \sim N(\mu,\sigma^2)$ where σ^2 is known.

Suppose that the critical values $c_1, c_2, \ldots, c_K$ define a feasible stopping rule for the hypothesis testing problem of §4.3, and that $c_1', c_2', \ldots, c_K'$ are the corresponding standardized critical values (i.e. $c_i' = c_i/\sqrt{in\sigma^2}$, $i = 1, 2, \ldots, K$). For $i = 1, 2, \ldots, K$, let $\bar{X}(i)$ denote the mean of the first i groups of observations (i.e. $\bar{X}(i) = S_{in}/in$), and, to ease notation, let

$$\underline{\mu}_i = \bar{X}(i) - \frac{\sigma}{\sqrt{in}} c_i'$$

and

$$\bar{\mu}_i = \bar{X}(i) + \frac{\sigma}{\sqrt{in}} c_i',$$

it follows that

$$\Pr(\underline{\mu}_i < \mu < \bar{\mu}_i \text{ for all } 1 \le i \le K) = (1-\alpha). \qquad (4.12.1)$$

We term $\{\underline{\mu}_i, \bar{\mu}_i);\ i = 1, 2, \ldots, K\}$ $(1-\alpha)$ RCIs for the parameter μ. As has already been mentioned these RCIs are valid no matter how a decision to stop a study is made. We can, however, recover our original group sequential hypothesis test from the RCIs. This is achieved by adopting the following stopping rule :-

At analysis i $(1 \le i \le K-1)$,

if $\underline{\mu}_i > 0$ stop entering patients on to the trial and accept H_1^+;

if $\bar{\mu}_i < 0$ stop entering patients on to the trial and accept H_1^-;

if $\underline{\mu}_i < 0 < \bar{\mu}_i$ enter the next group of $2n$ patients on to the trial.

At analysis K,

if $\underline{\mu}_K > 0$ stop entering patients on to the trial and accept H_1^+;

if $\bar{\mu}_K < 0$ stop entering patients on to the trial and accept H_1^-;

if $\underline{\mu}_K < 0 < \bar{\mu}_K$ stop entering patients on to the trial and accept H_0.

Jennison & Turnbull (1984), (1989) based their RCIs on the critical values for the tests of Pocock (1977) and O'Brien & Fleming (1979). Clearly by basing RCIs on the critical values of one of our optimal group sequential tests we obtain a procedure which is both flexible and highly efficient. We do not consider any results here, as the expected sample sizes of these RCIs are identical to the expected sample sizes of our optimal tests

The RCI approach can also be used when testing an experimental treatment against a standard as was the case in §3. We shall describe the method for the symmetric problem considered in §3.5. The generalization of this symmetric problem to the asymmetric problem, considered by Freedman & Spiegelhalter (1983) and Freedman, Lowe & Macaskill (1984), is quite straightforward.

Consider the hypothesis testing problem described in §3.4 with a maximum of K groups of n pairs of patients for testing $H_1^-: \mu = -\delta$ *vs* $H_1^+: \mu = \delta$ $(\delta > 0)$.

Suppose that the set of standardized critical values $c'_1, c'_2, \ldots, c'_K$ defines a feasible stopping rule for this problem. Using the same notation as before, it follows that

$$\Pr(\underline{\mu}_i < \mu < \bar{\mu}_i \text{ for all } 1 \le i \le K) = (1-2\alpha).$$

The intervals $\{\underline{\mu}_i, \bar{\mu}_i);\ i = 1, 2, \ldots, K\}$ are then $1-2\alpha$ RCIs for the parameter μ.

We can recover our one-sided test by adopting the following stopping rule:-

At analysis i $(1 \le i \le K-1)$,

if $\underline{\mu}_i > -\delta$ stop entering patients on to the trial and accept H_1^+;

if $\bar{\mu}_i < \delta$ stop entering patients on to the trial and accept H_1^-;

if $\underline{\mu}_i < -\delta$ and $\bar{\mu}_i > \delta$ enter the next group of n pairs of patients on to the trial.

At analysis K,

if $\underline{\mu}_K > -\delta$ stop entering patients on to the trial and accept H_1^+;

if $\bar{\mu}_K < \delta$ stop entering patients on to the trial and accept H_1^-.

Choosing the width of the Kth RCI less than 2δ ensures that the trial is terminated at the Kth analysis. This is equivalent to choosing

$$n > \frac{\sigma^2 c_K'}{\delta^2 K}.$$

Again, by basing these RCIs on a stopping rule for one of the optimal tests of §3 we obtain a procedure which is both flexible and highly efficient. Results for the expected sample sizes of these RCIs are not included here for the same reasons as cited earlier.

4.13 Discussion and Conclusions.

In §4 we have considered an improved method for the computation of optimal two-sided group sequential tests on the mean of a normal distribution with known variance. Our improved method is a generalization of that used for computing optimal one-sided tests (described in §3). The improved method is both computationally efficient and numerically stable.

The resulting tests are highly efficient and a substantial improvement on the fixed sample size test. They are also improvements (in some cases substantial improvements) on the two-sided group sequential tests proposed in the literature.

By basing the repeated confidence interval approach of Jennison & Turnbull (1984), (1989) on the stopping rule of one of our optimal two-sided tests, we obtain a procedure which is both flexible and optimal.

We have not extended our method to problems with unequal group sizes. Such an extension would be quite simple however. Likewise we could easily consider the minimization of objective functions over group sizes.

5. Optimal Inner Wedge Tests.

5.1 Introduction.

In §4 we considered two-sided sequential tests which allowed for the early rejection of the null hypothesis. Acceptance of H_0 was only permitted at the final analysis. Gould & Pecore (1982), Gould (1983), Whitehead & Stratton (1983) and Emerson & Fleming (1989) have all proposed two-sided tests which permit the early acceptance of H_0. We shall term such tests **inner wedge tests** or, simply, **wedge tests.**

As an example of where a wedge test might be appropriate, consider testing two experimental treatments, A and B, say. If A and B are therapeutically equivalent we would obviously like to stop our clinical trial as soon as possible in order to save money and to release valuable resources for use on other studies. Wedge tests are particularly appropriate for clinical trials designed to test for the bioequivalence of two drugs. We shall consider using optimal wedge tests on bioequivalence studies later on in this Chapter.

We shall give a formal description of a wedge test in §5.2 and review the literature on this topic in §5.3. In §5.4 we introduce optimal wedge tests. By generalizing our method for computing optimal two-sided tests (described in §§4.6, 4.7 and 4.8) we obtain an efficient and stable method for computing optimal wedge tests. In §§5.5, 5.6 and 5.7 we describe our method before giving some results in §§5.8 and 5.9. A comparison of our optimal wedge tests with some of the optimal two-sided tests of §4 and other wedge tests proposed in the literature is given in §5.10. In §5.11 we consider the use of optimal wedge tests in the bioequivalence problem.

5.2 Inner Wedge Tests.

Consider a clinical trial with a maximum of K groups of n pairs of patients available for testing the relative efficacies of two experimental treatments, A and B, say. Let X_i be the random variable representing the difference in response between the ith patient on treatment A and the ith patient on treatment B.

Suppose the X_i's are independent and normally distributed with unknown mean μ and known variance σ^2, and we wish to test

$$H_1^-: \mu < 0 \quad vs \quad H_0: \mu = 0 \quad vs \quad H_1^+: \mu > 0$$

with error rates

$$\Pr(\mathcal{A}_1^- \cup \mathcal{A}_1^+ | \mu = 0) = \alpha \tag{5.2.1}$$

$$\Pr(\mathcal{A}_0 \cup \mathcal{A}_1^+ | \mu = -\delta) = \beta \tag{5.2.2}$$

$$\Pr(\mathcal{A}_1^- \cup \mathcal{A}_0 | \mu = \delta) = \beta. \tag{5.2.3}$$

Here, as in §4, $\mathcal{A}_1^-$, $\mathcal{A}_0$ and $\mathcal{A}_1^+$ denote the acceptance of H_1^-, H_0 and H_1^+ respectively.

Also as in §4, the fixed sample size test for this problem requires N_f patients on each treatment arm, where N_f is given by equation (5.2.4)

$$N_f = \frac{\sigma^2}{\delta^2}\{\Phi^{-1}(1-\beta) + \Phi^{-1}(1-\alpha/2)\}^2. \tag{5.2.4}$$

For the group sequential wedge test we shall consider stopping rules of the general form:-

At analysis i $(1 \le i \le K-1)$,

if $S_{in} \ge c_i$ stop entering patients on to the trial and accept H_1^+;
if $S_{in} \le -c_i$ stop entering patients on to the trial and accept H_1^-;
if $|S_{in}| \le l_i$ stop entering patients on to the trial and accept H_0;
if $l_i < |S_{in}| < c_i$ enter the next group of n pairs of patients on to the trial.

At analysis K,

if $S_{Kn} \ge c_K$ stop entering patients on to the trial and accept H_1^+;
if $S_{Kn} \le -c_K$ stop entering patients on to the trial and accept H_1^-;
if $|S_{Kn}| \le c_K$ stop entering patients on to the trial and accept H_0 .

Here we require $l_i \le c_i$ $(1 \le i \le K-1)$.

We term a set of critical values $\{(l_1, c_1), \ldots, (l_{K-1}, c_{K-1}), c_K\}$ **feasible** if it defines a test satisfying the error constraints (5.2.1) - (5.2.3). Any test with a

feasible stopping rule is termed a **feasible test.** Clearly, for a given, sensibly formulated, problem there will exist infinitely many feasible tests.

As with the sequential tests of §§3 and 4 the sample size of a sequential wedge test is a random variable. The expected sample size under a given treatment difference μ, $E(N|\mu)$, is given by equation (5.2.5)

$$E(N|\mu) = n\sum_{j=1}^{K} j \int_{s_j}\int_{r_{j-1}} \dots \int_{r_1} f_\mu(x_1)\, f_\mu(x_2-x_1)\dots f_\mu(x_j-x_{j-1})\, dx_1 \dots dx_{j-1}\, dx_j \qquad (5.2.5)$$

where $f_\mu(x)$ is a normal density with mean $n\mu$ and variance $n\sigma^2$, $r_i = \{(-c_i,-l_i)\cup(l_i,c_i)\}$ for $i=1,2,\dots,K-1$ and $s_i = \{(-\infty,-c_j]\cup[-l_j,l_j]\cup[c_j,\infty)\}$ for $j=1,2,\dots,K$. We can calculate $E(N|\mu)$ numerically (details are given in Appendix 5.1).

The maximum sample size for our problem is nK observations (or $2nK$ patients). Wedge tests are compared in terms of their expected and maximum sample sizes in §5.10.

5.3 A Review of the Literature on Sequential Wedge Tests.

Sobel & Wald (1949) were the first to propose an inner wedge test. They suggested using two one-sided sequential probability ratio tests (see §3.4 for a description of the SPRT): one for testing H_1^-: $\mu = -\delta$ *vs* H_0: $\mu = 0$ and the other for testing H_0: $\mu = 0$ *vs* H_1^+: $\mu = \delta$. The test is fully sequential (i.e. $n = 1$) and open (i.e. there is no upper bound on the maximum sample size of the test). The critical values of the first test are given by

$$c_i' = -\frac{i\delta}{2} + \frac{\sigma^2}{\delta}\ln\left[\frac{\beta}{1-\alpha}\right] \quad \text{and} \quad l_i' = -\frac{i\delta}{2} + \frac{\sigma^2}{\delta}\ln\left[\frac{1-\beta}{\alpha}\right].$$

while the critical values of the second test are given by

$$l_i = \frac{i\delta}{2} + \frac{\sigma^2}{\delta}\ln\left[\frac{\beta}{1-\alpha}\right] \quad \text{and} \quad c_i = \frac{i\delta}{2} + \frac{\sigma^2}{\delta}\ln\left[\frac{1-\beta}{\alpha}\right]$$

Sampling stops when **both** tests have terminated, with the acceptance of H_1^+ if S_i $(= X_1 + \dots + X_i) \geq c_i$ and the acceptance of H_1^- if $S_i \leq c_i'$.

The continuation regions of the two tests overlap for $i < i'$, where

$$i' = \frac{\sigma^2}{\delta^2}\left\{\ln\left[\frac{\beta}{1-\alpha}\right] - \ln\left[\frac{1-\beta}{\alpha}\right]\right\}.$$

In order to preserve the overall error rates of the test H_0 is accepted if **both** l_i and l_i' are crossed before $i = i'$. For $i \geq i'$ H_0 is accepted if $l_i' \leq S_i \leq l_i$.

Unfortunately the open nature of this test together with the requirement that data monitoring is fully sequential make it unsuitable for use in clinical trials.

In order to place an upper bound on the maximum sample size, Schneiderman & Armitage (1962) and Armitage (1975, Ch. 5) proposed a closed or restricted sequential wedge test with a maximum of K observations. They pointed out that for the two-sided restricted sequential test of Armitage (1957) (described in § 4.4) the probability of rejecting H_0 is small when the number of observations, i, is close to K and $S_i \approx 0$. Hence it would seem reasonable to generalize this test to allow for the early acceptance of H_0. Obviously there are infinitely many ways of defining an inner wedge for this test. Schneiderman & Armitage suggested using the locus of points ζ where the null probability of crossing either of the outer boundaries from any point on ζ equals some small constant, ε'. In order to preserve the overall size of the test Schneiderman & Armitage recommended increasing K to K'. (Alternatively we could fix K and widen the boundaries of the continuation region). The Type II error of the test, β, will, in general, change but, typically, not by any substantial amount. The resulting procedure leads to large savings in $E(N|\mu)$ compared with the original two-sided restricted test when $|\mu|$ is small.

Gould & Pecore (1982) and Gould (1983) suggested modifying the two-sided group sequential test of Pocock (1977) (described in §4.4) to allow for the early acceptance of H_0. They set $c_i = c\sqrt{i}$ $(i = 1, 2, \ldots, K)$ and $l_i = l\sqrt{i}$ $(i = 1, 2, \ldots, K-1)$. The experimenter is free to choose l while c is constrained to give a test of size α. Clearly c and l are related; for $K = 2$ and $\alpha = 0.05$ Gould & Pecore (1982) claimed that a good polynomial approximation to this relationship is given by

$$c = 2.173 - 0.0124l + 0.0005l^2 - 0.27l^3 .$$

Although neither paper talks explicitly about controlling the Type II error rate it is clear that a suitable choice of group size gives a test with errors β at $\mu = \pm\delta$.

Pocock's test is a special case of the Gould & Pecore test with $l = 0$. For n and K fixed, the Pocock test maximizes the power of detecting the treatment

difference $\mu = \pm\delta$ over the family of Gould & Pecore tests. Unfortunately it also maximizes the expected sample size under $\mu = 0$ over the same family. Increasing l (with n and K still fixed) reduces both the power of detecting $\mu = \pm\delta$ and $E(N|\mu=0)$.

Gould & Pecore pointed to the robustness of their tests to departures from normality. With $K = 2$, π_1 $(= 2\Phi(-l))$ equal to 0.5 and 1.0, $\alpha = 0.05$ and $\beta = 0.1$ and 0.2, Monte Carlo simulations were conducted on data from the binomial and Student t distributions. The observed error rates were impressively close to α and β. Further simulations were conducted on the Student t distribution, this time with unpredictable group sizes. Again the error rates of the tests were close to α and β.

A major criticism of the Gould & Pecore test concerns the form of its inner wedge at the first few analyses when accrued sample sizes tend to be small. Clearly it is dubious, to say the least, to interpret a small test statistic at, say, the first analysis, as strong evidence in favour of treatment equivalence. One possible way of overcoming this problem would be to set $l_i = 0$ for $i \leq I$ and $l_i = l\sqrt{i}$ for $I+1 \leq i \leq K-1$. The choice of I being made at the design stage of the experiment.

Whitehead (1983, Ch. 4) and Whitehead & Stratton (1983) suggested using two one-sided triangular tests for the wedge problem: one for testing H_1^-: $\mu < 0$ *vs* H_0: $\mu = 0$ and the other for testing H_0: $\mu = 0$ *vs* H_1^+: $\mu > 0$. As was mentioned in §3.3 the triangular test assumes continuous data monitoring and it is necessary to adjust the boundaries to take into account the effect of discrete monitoring. Therefore the critical values for the first test are given by

$$c_i' = \frac{-2}{\delta}\log\left[\frac{1}{\alpha+\beta}\right] + 0.583\sqrt{n} - \frac{\delta}{4}ni$$

and

$$l_i' = \frac{2}{\delta}\log\left[\frac{1}{\alpha+\beta}\right] - 0.583\sqrt{n} - \frac{3\delta}{4}ni,$$

while the critical values for the second test are given by

$$l_i = \frac{-2}{\delta}\log\left[\frac{1}{\alpha+\beta}\right] + 0.583\sqrt{n} + \frac{3\delta}{4}ni$$

and

$$c_i = \frac{2}{\delta}\log\left[\frac{1}{\alpha+\beta}\right] - 0.583\sqrt{n} + \frac{\delta}{4}ni$$

where the constant 0.583 is a correction factor for discreteness due to Siegmund (1979) and Cuzick (1981). As with the test of Sobel & Wald (1949) sampling continues until **both** tests have terminated, with the acceptance of H_1^+ if $S_{in} \geq c_i$, and the acceptance of H_1^- if $S_{in} \leq c_i'$. The continuation regions of the tests overlap for

$$ni < \frac{8}{3\delta^2}\log\left[\frac{1}{\alpha+\beta}\right] - \frac{0.583\text{x}4\sqrt{n}}{3\delta} = t^*$$

To preserve the overall error rates of the test we accept H_0 if **both** l_i and l_i' are crossed before $ni = t^*$. For $ni \geq t^*$, H_0 is accepted if $l_i' \leq S_{in} \leq l_i$.

Whitehead & Stratton compared their test with that of Pocock (1977), by considering an example and some simulations. In particular they pointed to the large savings in $E(N|\mu=0)$ made possible by their test.

Emerson & Fleming (1989) proposed a family of symmetric wedge tests where $\beta = \alpha/2$. The critical values of these tests are given by $c_i = (ni)^p z$ and $l_i = \max(0, ni\delta - c_i)$, where p is a design parameter, z is chosen to give a test of size α and the group size, n, is constrained so that $l_K = c_K$, i.e.

$$n = \frac{1}{K^{p-1}}\left[\frac{\delta}{2z}\right]^{\frac{1}{p-1}}.$$

The above choice of group size means that, in general, the Type II error constraint will not be satisfied exactly. Emerson & Fleming stated that '... The power under the alternative hypothesis for a level .05 two-sided symmetric test is typically .976 when $p = 0$ and .98 when $p = .5$.'

For $K \leq 10$ and $\alpha = 0.01$ and 0.05, Emerson & Fleming compared their tests for $p = 0$ and $p = 0.5$ with the two-sided tests of Pocock (1977) and O'Brien & Fleming (1979), and the wedge test of Whitehead & Stratton (1983), in terms of both $F_1 : E(N|\mu=0)$ and $F_2 : E(N|\mu=\delta)$. For objective function F_1 the three wedge tests exhibit large gains in efficiency compared with the two-sided tests. Under $\mu = \delta$ the Pocock test leads to the smallest expected sample sizes with the Emerson & Fleming test with $p = 0.5$ almost as efficient.

5.4 Optimal Group Sequential Wedge Tests.

For a given problem, with n, K, α, β and δ fixed, the optimal wedge test minimizes some chosen objective function, F, over the set of feasible wedge tests.

In this chapter we shall consider the minimization of the following 3 objective functions:

$$F_1 : E(N|\mu=0),$$

$$F_2 : E(N|\mu=\delta)$$

and

$$F_5 : \int E(N|\mu)\, \delta^{-1}\, \varphi(\mu/\delta)\, d\mu.$$

Of particular interest is the minimization of F_1 now that the early acceptance of H_0 is permitted. In §5.8 we give results for the minimization of F_1, while in §5.10 we compare the optimal wedge tests for F_1 with some of the other tests proposed in the literature.

The computation of optimal wedge tests by a direct numerical search over feasible stopping rules is, in general, impractical because of the large number of unconstrained critical values. For instance with K as small as 5 there are 7 free boundary points to optimize over, and the simplex algorithm of Nelder & Mead (1965) would begin to experience problems (see §3.5 for a more detailed discussion of the inadequacies of the Nelder & Mead algorithm). Instead we choose to extend our improved method for the computation of optimal two-sided tests to wedge tests. Again, unlike the direct numerical search method, our improved approach is computationally efficient and numerically stable. In §5.5 we describe our improved method for objective function $F_1 : E(N|\mu=0)$, while in §5.6 we indicate how to adapt our approach when considering other objective functions.

5.5 An Improved Method for the Computation of Optimal Wedge Tests.

Consider again the clinical trials problem described in §5.2 with a maximum of K equally sized groups of n pairs of patients. The random variable X_i $(i = 1, 2, \ldots, nK)$ denotes the difference in response between the ith pair of patients and is normally distributed with unknown mean μ and known variance

σ^2.

Here we wish to make a choice between the three decisions :

$$D_1^-: \mu < 0, \qquad D_0: \mu = 0 \qquad \text{and } D_1^+: \mu > 0.$$

We shall consider a family of Bayes decision theory problems with a common prior distribution for μ given by $\pi(-\delta) = \pi(0) = \pi(\delta) = 1/3$ with $\pi(\mu) = 0$ otherwise, and a common cost of sampling function given by $c(0) = 1$ with $c(\mu) = 0$ otherwise. Individual problems within the family differ in their loss functions, $L(D,\mu)$, which are indexed by a pair of parameters, d_0 (>0) and d_1 (>0). The general form of $L(D,\mu)$ is given by:

$$L(D_0,-\delta) = L(D_1^+,-\delta) = d_1$$

$$L(D_1^-,0) = L(D_1^+,0) = d_0$$

$$L(D_1^-,\delta) = L(D_0,\delta) = d_1$$

with $L(D,\mu) = 0$ otherwise.

Suppose for the moment that d_0 and d_1 are fixed. Consider a general decision rule for the above problem which we shall denote by $\mathcal{B}$. Because our family of Bayes decision problems is symmetric about $\mu=0$ we shall only consider decision rules which are symmetric about zero. So $\mathcal{B}$ is of the general form :

At analysis i $(1 \le i \le K-1)$,

if $S_{in} \ge c_i$ stop entering patients on to the trial and make decision D_1^+;

if $S_{in} \le -c_i$ stop entering patients on to the trial and make decision D_1^-;

if $|S_{in}| \le l_i$ stop entering patients on to the trial and make decision D_0;

if $l_i < |S_{in}| < c_i$ enter the next group of n pairs of patients on to the trial.

At analysis K,

if $S_{Kn} \ge c_K$ stop entering patients on to the trial and make decision D_1^+;

if $S_{Kn} \le -c_K$ stop entering patients on to the trial and make decision D_1^-;

if $|S_{Kn}| \le c_K$ stop entering patients on to the trial and make decision D_0 .

Here we require $l_i \le c_i$ $(1 \le i \le K-1)$.

The **risk** associated with $\mathcal{B}$, which we shall denote by $r(\mathcal{B},d_0,d_1)$, is defined as the sum of the total expected sampling cost of the trial plus the total expected loss through making a wrong decision. That is

$$r(\mathcal{B},d_0,d_1) = c(0)\,E(N|\mu=0)\,\pi(0) + d_1\,\Pr(D_0\cup D_1^{+}|\mu=-\delta)\,\pi(-\delta)$$
$$+ d_0\,\Pr(D_1^{-}\cup D_1^{+}|\mu=0)\,\pi(0) + d_1\,\Pr(D_1^{-}\cup D_0|\mu=\delta)\,\pi(\delta)$$

where $\Pr(D_i\cup D_j|\mu)$ denotes the probability of making either decision D_i or decision D_j under a treatment difference μ.

It follows that

$$r(\mathcal{B},d_0,d_1) = \frac{1}{3}\{F_1 + d_1\,\Pr(D_0\cup D_1^{+}|\mu=-\delta) + d_0\,\Pr(D_1^{-}\cup D_1^{+}|\mu=0)$$
$$+ d_1\,\Pr(D_1^{-}\cup D_0|\mu=\delta)\}.$$

The Bayes decision rule for our problem, $\mathcal{B}^*(d_0,d_1)$, minimizes this risk over the set of all decision rules, $\mathcal{S}$, i.e.

$$r(\mathcal{B}^*(d_0,d_1),d_0,d_1) = \min_{\mathcal{B}\in\mathcal{S}}\{r(\mathcal{B},d_0,d_1)\}.$$

We can compute $\mathcal{B}^*(d_0,d_1)$ by dynamic programming (the necessary computations are described in §5.7). Using numerical integration we can calculate the Type I and Type II errors of $\mathcal{B}^*(d_0,d_1)$. Suppose these errors are given by

$$\Pr(D_1^{-}\cup D_1^{+}|\mu=0) = \tilde{\alpha}$$

and

$$\Pr(D_0\cup D_1^{+}|\mu=-\delta) = \Pr(D_1^{-}\cup D_0|\mu=\delta) = \tilde{\beta},$$

then clearly

$$r(\mathcal{B}^*(d_0,d_1),d_0,d_1) = \frac{1}{3}\{F_1 + 2\,d_1\,\tilde{\beta} + d_0\,\tilde{\alpha}\}.$$

It follows from the definition of the Bayes rule that there can be no other decision rule with errors $\tilde{\alpha}$ and $\tilde{\beta}$ which attains a lower value of F_1.

Using the algorithm of Powell (1970) we search over (d_0,d_1) for a pair of loss parameters $(d_0^{(\alpha)},d_1^{(\beta)})$ giving rise to a Bayes decision theory problem with an associated Bayes rule, $\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)})$, which has Type I error α and Type II error β. Clearly

$$r(\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)}),d_0^{(\alpha)},d_1^{(\beta)}) = \frac{1}{3}\{F_1 + 2\,d_1^{(\beta)}\,\beta + d_0^{(\alpha)}\,\alpha\}.$$

Again, from the definition of the Bayes rule, there can be no other decision rule for this problem with a smaller risk, i.e.

$$r(\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)}),d_0^{(\alpha)},d_1^{(\beta)}) = \min_{\mathcal{B}\in\mathcal{S}} \{r(\mathcal{B},d_0^{(\alpha)},d_1^{(\beta)})\}.$$

Moreover there can be no other decision rule with errors α and β which attains a lower value of F_1. By equating decisions D_1^-, D_0 and D_1^+ with the acceptance of the hypotheses H_1^-, H_0 and H_1^+ respectively we have computed the optimal two-sided group sequential test for our original frequentist problem.

5.6 Applying our Improved Method to Other Objective Functions.

Our improved method for the computation of optimal wedge tests described in §5.5 is easily adapted for use with other objective functions. The logic of the method has been described a number of times already and will be omitted here. We will, however, describe suitable families of Bayes decision theory problems for the minimization of objective functions $F_2: E(N|\mu=\delta)$ and $F_5: \int E(N|\mu)\,\delta^{-1}\,\varphi(\mu/\delta)\,d\mu$.

Consider the minimization of F_2. A suitable family of Bayes decision theory problems has the common prior distribution for μ given by : $\pi(-\delta) = \pi(0) = \pi(\delta) = 1/3$ with $\pi(\mu) = 0$ otherwise, and a common cost of sampling function given by : $c(-\delta) = c(\delta) = 1$ with $c(\mu) = 0$ otherwise. The general form of the loss function, $L(D,\mu)$, is identical to that given in §5.5, with, in particular, $L(D,\mu) = 0$ for $\mu \neq 0, \pm\delta$. The rest of the logic is identical to that given in §4.6.

For the minimization of F_5, a suitable family of Bayes decision theory problems has the common prior distribution for μ given by :

$$\pi(\mu) = \begin{cases} 1/4 & \text{if } \mu = -\delta, 0 \text{ or } \delta \\ (1/4)\delta^{-1}\varphi(\mu/\delta) & \text{otherwise,} \end{cases}$$

and the common cost of sampling function :

$$c(\mu) = \begin{cases} 1 \text{ if } \mu \neq -\delta, 0 \text{ or } +\delta \\ 0 \text{ otherwise.} \end{cases}$$

The general form of the loss function is identical to that given in §5.5, with, in

particular, $L(D,\mu) = 0$ for $\mu \neq -\delta, 0, \delta$.

The rest of the logic is identical to that given in §4.7 when considering the minimization of F_5 over feasible two-sided group sequential tests.

5.7 The Dynamic Programming Algorithm.

As in §§3 and 4, for any given Bayes decision theory problem the Bayes rule is computed using dynamic programming. Here we shall describe the dynamic programming algorithm for a general wedge problem. Throughout we shall assume that the Bayes rule is monotone. That is, at analysis i $(i = 1, 2, \ldots, K)$ it is optimal to make decision ${D_1}^+$ if $S_{in} \geq c_i$ and decision ${D_1}^-$ if $S_{in} \leq -c_i$ and decision D_0 if $-l_i \leq S_{in} \leq l_i$, while for $l_i < |S_{in}| < c_i$ it is optimal to sample the next group of n observations. Numerical checks support our assumption that $\mathcal{B}^*(d_0, d_1)$ is monotone.

Consider again the Bayes decision theory problems of §5.5 with a maximum of K groups of n pairs of patients available for choosing between the three decisions

$$ {D_1}^-: \mu < 0, \quad D_0: \mu = 0 \quad \text{and} \quad {D_1}^+: \mu > 0. $$

This time suppose that we have a prior distribution which is denoted by $\pi(\mu)$ and is defined over some parameter space M, and that our cost of sampling function is denoted by $c(\mu)$ and is also defined over M. Finally we shall suppose that our loss function is denoted by $L(D,\mu)$ and that it is of the same general form as that given in §5.5.

Suppose the loss parameters d_0 and d_1 are fixed. Letting $p^{(i)}(\mu|x)$ denote the current posterior distribution of μ at analysis i $(1 \leq i \leq K)$, given that $S_{in} = x$, then the loss from making decision ${D_1}^+$ equals

$$ d_1\, p^{(i)}(-\delta|x) + d_0\, p^{(i)}(0|x), $$

the loss from making decision D_0 equals

$$ d_1\, \{p^{(i)}(-\delta|x) + p^{(i)}(\delta|x)\} $$

and the loss from making decision ${D_1}^-$ equals

$$ d_0\, p^{(i)}(0|x) + d_1\, p^{(i)}(\delta|x). $$

If we let $\gamma^{(i)}(x)$ denote the minimum loss from stopping at analysis i $(i = 1, 2, \ldots, K)$ and making a definite decision, it follows that

$$\gamma^{(i)}(x) = \min\big[\{d_0\, p^{(i)}(0|x) + d_1\, p^{(i)}(\delta|x)\}, d_1\,\{p^{(i)}(-\delta|x) + p^{(i)}(\delta|x)\},$$
$$\{d_1\, p^{(i)}(-\delta|x) + d_0\, p^{(i)}(0|x)\}\big]$$

Further, let $\beta^{(i)}(x)$ denote the minimum additional risk from sampling the $(i+1)$st group $(i = 1, 2, \ldots, K-1)$ and then proceeding optimally. Denoting the c.d.f. of $S_{(i+1)n}$ given that $S_{in} = x$ by $F^{(i+1)}(S_{(i+1)n}|x)$, it follows that

$$\beta^{(K-1)}(x) = n \sum_{\mu \in M} c(\mu) p^{(K-1)}(\mu|x) + \int_{S_{Kn}} \gamma^{(K)}(S_{Kn})\, dF^{(K)}(S_{Kn}|x)$$

and, for $i < K-1$,

$$\beta^{(i)}(x) = n \sum_{\mu \in M} c(\mu) p^{(i)}(\mu|x)$$
$$+ \int_{S_{(i+1)n}} \min\{\beta^{(i+1)}(S_{(i+1)n}), \gamma^{(i+1)}(S_{(i+1)n})\}\, dF^{(i+1)}(S_{(i+1)n}|x)$$

where the summation signs in the above equations are replaced by a mixture of summations and integrals for objective functions such as F_5.

We compute the critical values of $\mathcal{B}^*(d_0, d_1)$ by starting at the Kth analysis and working back.

At analysis K it is optimal to make decision ${D_1}^+$ for all x (≥ 0) such that

$$d_1\, p^{(K)}(-\delta|x) + d_0\, p^{(K)}(0|x) < d_1\,\{p^{(K)}(-\delta|x) + p^{(K)}(\delta|x)\} \quad (5.7.1)$$

and to make decision D_0 for all x (≥ 0) such that

$$d_1\,\{p^{(K)}(-\delta|x) + p^{(K)}(\delta|x)\} < d_1\, p^{(K)}(-\delta|x) + d_0\, p^{(K)}(0|x). \quad (5.7.2)$$

Using the bisection method we obtain the Kth critical value, c_K, as the solution, for $x \geq 0$, of the equation

$$d_1\,\{p^{(K)}(-\delta|x) + p^{(K)}(\delta|x)\} = d_1\, p^{(K)}(-\delta|x) + d_0\, p^{(K)}(0|x). \quad (5.7.3)$$

Clearly it is possible that equation (5.7.3) does not possess a solution for $x \geq 0$. To avoid our algorithm running in to computational difficulties we can build a simple check in to our program. The test checks whether inequality (5.7.2) holds for $x = 0$. If this is not the case we set $c_K = 0$ and go back to the $(K-1)$st analysis. A similar check can be built in at other analyses.

The symmetry inherent in our problem makes it optimal to choose decision D_1^- if $S_{Kn} \leq -c_K$, and decision D_0 if $-c_K < S_{Kn} \leq 0$.

At analysis $K-1$ it is optimal to make decision D_1^+ for all $S_{(K-1)n} = x \ (\geq 0)$ such that

$$d_1\, p^{(K-1)}(-\delta|x) + d_0\, p^{(K-1)}(0|x) < \beta^{(K-1)}(x)$$

and **not** to make decision D_1^+ for all $x \ (\geq 0)$ such that

$$\beta^{(K-1)}(x) < d_1\, p^{(K-1)}(-\delta|x) + d_0\, p^{(K-1)}(0|x),$$

where the function $\beta^{(K-1)}(x)$ is given by

$$\beta^{(K-1)}(x) = n \sum_{\mu \in M} c(\mu) p^{(K-1)}(\mu|x)$$

$$+ d_1 \Pr_{-\delta}(S_{Kn} \geq c_K \mid x)\, p^{(K-1)}(-\delta|x) + d_0 \Pr_0(S_{Kn} \geq c_K \mid x)\, p^{(K-1)}(0|x)$$

$$+ d_1 \{\Pr_{-\delta}(-l_K \leq S_{Kn} \leq l_K \mid x)\, p^{(K-1)}(-\delta|x) + \Pr_{\delta}(-l_K \leq S_{Kn} \leq l_K \mid x)\, p^{(K-1)}(\delta|x)\}$$

$$+ d_0 \Pr_0(S_{Kn} \leq -c_K \mid x)\, p^{(K-1)}(0|x) + d_1 \Pr_{\delta}(S_{Kn} \leq -c_K \mid x)\, p^{(K-1)}(\delta|x).$$

Again we use the bisection method to obtain c_{K-1} as the solution, for $x \geq 0$, of equation (5.7.4)

$$d_1\, p^{(K-1)}(-\delta|x) + d_0\, p^{(K-1)}(0|x) = \beta^{(K-1)}(x). \qquad (5.7.4)$$

The symmetry inherent in our problem makes it optimal to choose D_1^- if $S_{(K-1)n} \leq -c_{K-1}$ and **not** to choose D_1^- if $-c_{K-1} < S_{n(K-1)} \leq 0$.

Also at analysis $K-1$ it is optimal to make decision D_0 for $S_{(K-1)n} = x \ (\geq 0)$ such that

$$d_1 \{p^{(K-1)}(-\delta|x) + p^{(K-1)}(\delta|x)\} < \beta^{(K-1)}(x)$$

and **not** to make decision D_0 for all $x \ (\geq 0)$ such that

$$\beta^{(K-1)}(x) < d_1 \{p^{(K-1)}(-\delta|x) + p^{(K-1)}(\delta|x)\}.$$

We obtain l_{K-1} as the solution, for $x \geq 0$, of the equation

$$d_1 \{p^{(K-1)}(-\delta|x) + p^{(K-1)}(\delta|x)\} = \beta^{(K-1)}(x).$$

The symmetry inherent in our problem makes it optimal to choose D_0 if $-l_{K-1} \leq S_{n(K-1)} \leq 0$ and **not** to choose D_0 if $S_{n(K-1)} < -l_{K-1}$.

Clearly it is possible that $l_{K-1} > c_{K-1}$. If this is the case, sampling is terminated at the $(K-1)$st analysis with a decision in favour of $D_1{}^+$ if $S_{in} \geq c_i{}^*$ and a decision in favour of D_0 if $0 \leq S_{in} < c_i{}^*$, where $c_i{}^*$ is the solution, for $x \geq 0$, of the equation

$$d_1 \{p^{(K-1)}(-\delta|x) + p^{(K-1)}(\delta|x)\} = d_1 p^{(K-1)}(-\delta|x) + d_0 p^{(K-1)}(0|x).$$

At analysis $K-2$ it is optimal to make decision $D_1{}^+$ for all $S_{(K-2)n} = x$ (≥ 0) such that

$$d_1 p^{(K-2)}(-\delta|x) + d_0 p^{(K-2)}(0|x) < \beta^{(K-2)}(x)$$

and **not** to make decision $D_1{}^+$ for all x (≥ 0) such that

$$\beta^{(K-2)}(x) < d_1 p^{(K-2)}(-\delta|x) + d_0 p^{(K-2)}(0|x),$$

where the function $\beta^{(K-2)}(x)$ is given by

$$\beta^{(K-2)}(x) = n \sum_{\mu \in M} c(\mu) p^{(K-2)}(\mu|x)$$

$$+d_1 \Pr_{-\delta}(S_{(K-1)n} \geq c_{K-1}|x) p^{(K-2)}(-\delta|x) + d_0 \Pr_0(S_{(K-1)n} \geq c_{K-1}|x) p^{(K-2)}(0|x)$$

$$+d_1 \{\Pr_{-\delta}(|S_{(K-1)n}| \leq l_{K-1}|x) p^{(K-2)}(-\delta|x) + \Pr_{\delta}(|S_{(K-1)n}| \leq l_{K-1}|x) p^{(K-2)}(\delta|x)\}$$

$$+d_0 \Pr_0(S_{(K-1)n} \leq -c_{K-1}|x) p^{(K-2)}(0|x) + d_1 \Pr_{\delta}(S_{(K-1)n} \leq -c_{K-1}|x) p^{(K-2)}(\delta|x)\}$$

$$+ \int_{l_{K-1}}^{c_{K-1}} \beta^{(K-1)}(S_{(K-1)n}) dF^{(K-1)}(S_{(K-1)n}|x) + \int_{-c_{K-1}}^{-l_{K-1}} \beta^{(K-1)}(S_{(K-1)n}) dF^{(K-1)}(S_{(K-1)n}|x).$$

Again the bisection method gives the $(K-2)$nd critical value, c_{K-2}, as the solution, for $x \geq 0$, of equation (5.7.5)

$$d_1 p^{(K-2)}(-\delta|x) + d_0 p^{(K-2)}(0|x) = \beta^{(K-2)}(x). \qquad (5.7.5)$$

The symmetry inherent in our problem makes it optimal to choose $D_1{}^-$ if $S_{(K-2)n} \leq -c_{K-2}$ and **not** to choose $D_1{}^-$ if $-c_{K-2} < S_{(K-2)n} \leq 0$.

Also at analysis $K-2$ it is optimal to make decision D_0 for $S_{(K-2)n} = x$ (≥ 0) such that

$$d_1 \{p^{(K-2)}(-\delta|x) + p^{(K-2)}(\delta|x)\} \leq \beta^{(K-2)}(x)$$

and **not** to make decision D_0 for all x (≥ 0) such that

$$\beta^{(K-2)}(x) < d_1 \{p^{(K-2)}(-\delta|x) + p^{(K-2)}(\delta|x)\}.$$

We obtain l_{K-2} as the solution, for $x \geq 0$, of the equation

$$d_1 \{p^{(K-2)}(-\delta|x) + p^{(K-2)}(\delta|x)\} = \beta^{(K-2)}(x).$$

The symmetrical nature of our problem makes it optimal to choose decision D_0 if $-l_{K-2} \leq S_{(K-2)n} \leq 0$ and **not** to choose D_0 if $S_{(K-2)n} < -l_{K-2}$.

We work back to the first analysis in a similar fashion. The error probabilities of the resulting decision rule are then computed and a test for convergence conducted. If convergence has not been achieved the error probabilities are fed into Powell's method and a new pair (d_0, d_1) obtained. The method is computationally very fast and efficient.

5.8 Results and Discussion.

Tables 5.1, 5.2 and 5.3 give the minima of $F_1: E(N|\mu=0)$, $F_2: E(N|\mu=\delta)$ and $F_5: \int E(N|\mu)\,\delta^{-1}\,\varphi(\mu/\delta)\,d\mu$ respectively expressed as percentages of the fixed sample size, N_f, for $\alpha = 0.05$, $\beta = 0.1$, K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200, and t, the ratio of the maximum sample size of the sequential test (nK) to N_f, equal to 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6. The minima are independent of δ and σ^2 (for verification of this fact see Appendix 5.2).

It is important to realize that the results given here are generally much better, and never any worse, than the results of §4.9. The reason for this is that, for any given problem, the set of feasible two-sided tests is a subset of the set of feasible wedge tests.

Table 5.1. Minima of F_1 : $E(N|\mu=0)$ expressed as percentages of the fixed sample size, N_f, for $\alpha=0.05$, $\beta=0.1$, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	92.1	79.9	76.2	**75.4**	75.6	77.3	79.6	82.4	85.5
3	82.7	77.9	75.0	73.0	71.9	**71.1**	71.1	71.6	72.4
4	82.2	73.5	71.3	70.9	70.6	70.1	69.9	**69.8**	69.8
5	79.6	73.0	69.5	68.2	**67.8**	68.1	68.6	68.9	69.1
10	77.2	70.0	66.8	65.6	64.9	**64.0**	64.0	64.2	64.3
15	76.3	69.0	66.0	64.6	63.8	63.1	62.8	**62.7**	62.8
20	75.8	68.5	65.5	64.0	63.3	62.5	62.1	**62.1**	62.2
30	75.4	68.1	65.1	63.6	62.8	62.0	61.6	61.5	**61.4**
50	75.0	67.7	64.7	63.2	62.4	61.6	61.2	61.0	**60.8**
100	74.7	67.4	64.4	62.9	62.1	61.2	60.8	60.6	**60.5**
200	74.6	67.3	64.3	62.8	62.0	61.1	60.6	60.4	**60.3**

Many of the comments made concerning the results of §§3 and 4 apply equally here. Again the largest gains in efficiency come from going from a single sample test to a sequential test with 2 groups. For fixed t, gains in efficiency decrease as K increases. Indeed an expected gain of no more than 4% of the fixed sample size is obtained in going from a 10 to a 200 group design. Even if economic considerations limit us to designs with $t=1.01$, Table 5.1 shows that substantial savings in $E(N|\mu=0)$ are possible.

If we compare the results of Table 5.1 for tests with $t=1.01$ and $K\leq 10$, with the corresponding results for two-sided tests given in Table 4.8, we see the very substantial gains which can be made from employing optimal wedge tests. Indeed, for any two-sided test with $\alpha=0.05$, F_1 must be at least 95% of the maximum sample size. The optimal wedge tests show substantial improvements on this lower bound for all the designs considered in Table 5.1.

One unique feature of Table 5.1, which is worth commenting on, is that the minimum values of F_1 for $K=3$ and $K=4$ occur at higher values of t than do the same minima for $K=5$ and $K=10$. The reasons for this are not clear, however they must be linked in some way with the introduction of the inner wedge.

Table 5.2. Minima of F_2: $E(N|\mu=\delta)$ expressed as percentages of the fixed sample size, N_f, for α=0.05, β=0.1, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	83.9	78.4	**77.2**	77.3	77.8	79.6	81.9	84.5	87.3
3	79.2	73.7	71.8	71.2	**71.1**	71.3	71.9	72.9	74.1
4	77.6	71.4	69.4	68.5	68.2	**68.1**	68.4	68.8	69.3
5	76.2	70.0	67.8	67.0	66.5	**66.2**	66.4	66.7	67.0
10	73.4	67.1	64.8	63.7	63.1	62.6	**62.4**	62.4	62.5
15	72.4	66.1	63.7	62.6	62.0	61.3	61.0	**60.9**	60.9
20	72.0	65.6	63.2	62.1	61.4	60.7	60.4	**60.2**	60.2
30	71.5	65.2	62.7	61.6	60.9	60.1	59.8	59.5	**59.4**
50	71.1	64.8	62.3	61.2	60.5	59.7	59.3	59.0	**58.9**
100	70.9	64.5	62.0	60.9	60.1	59.3	58.9	58.6	**58.4**
200	70.8	64.3	61.9	60.8	60.0	59.1	58.7	58.4	**58.2**

Table 5.2 exhibits many similar features to Table 5.1, with large gains in efficiency resulting from employing a 2 group design and only small gains resulting from using more than 10 groups. For $K \geq 4$ the minima of Table 5.1 are uniformly larger than the minima given here.

It is of interest to compare the results of Table 5.2 with those for designs without an inner wedge given in Table 4.2. For K large and/or t small the gains in efficiency for the wedge tests are quite small. With K = 2 and t = 1.6 a further saving of 6.7% of fixed sample size is possible on adding the inner wedge,

while for $K = 5$ and $t = 1.6$ this further saving is 6.9%. For K fixed the minimum of F_2 over t tends to occur at a higher maximum sample size for the wedge designs.

Table 5.3. Minima of F_5: $\int E(N|\mu)\,\delta^{-1}\,\varphi(\mu/\delta)\,d\mu$ expressed as percentages of the fixed sample size, N_f, for α=0.05, β=0.1, t = 1.01, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5 and 1.6 and K = 2, 3, 4, 5, 10, 15, 20, 30, 50, 100 and 200.

	t								
K	1.01	1.05	1.1	1.15	1.2	1.3	1.4	1.5	1.6
2	86.3	80.0	77.9	**77.5**	77.8	79.4	81.8	84.5	87.5
3	79.9	74.8	73.3	72.4	71.8	**71.5**	71.9	72.8	74.0
4	78.1	72.2	69.7	68.8	68.7	**68.6**	68.8	69.1	69.6
5	76.7	70.9	68.5	67.1	66.4	**66.1**	66.5	67.0	67.5
10	73.9	68.0	65.4	64.0	63.3	62.5	**62.1**	62.1	62.4
15	72.9	67.0	64.4	63.0	62.2	61.3	60.9	60.8	**60.7**
20	72.5	66.5	63.9	62.5	61.6	60.7	60.3	60.1	**60.0**
30	72.1	66.1	63.4	62.0	61.1	60.1	59.6	59.4	**59.3**
50	71.7	65.7	63.0	61.6	60.7	59.6	59.1	58.9	**58.8**
100	71.4	65.4	62.7	61.3	60.4	59.3	58.8	58.5	**58.4**
200	71.3	65.3	62.6	61.2	60.3	59.1	58.6	58.3	**58.2**

The main reason for computing optimal designs for objective function F_5 is to obtain tests which are optimal, or close to optimal, over the entire parameter space for μ. As can be seen from Table 5.3 the resulting tests are highly efficient. The table shares many of the characteristics of Tables 5.1 and 5.2.

As with objective functions F_1 and F_2 it is interesting to compare the results of Table 5.3 with the analogous ones for the two-sided test given in Table 4.6. The gains in efficiency from adopting a wedge design can be seen to be highly impressive. For example with $K = 3$ and $t = 1.6$ a further saving of 34.9% of the fixed sample size is achieved, while with $K = 10$ and $t = 1.5$ this further saving

is 33.7%. For t small and/or K large these savings are less impressive but still significant.

The minimization of objective function $F_3: E(N|\mu=2\delta)$ has not been considered in this Chapter. The expected gains in efficiency from considering a wedge test rather than a two-sided test are small and often non-existent here.

Clearly our improved method for computing optimal wedge tests could be extended to consider other objective functions and/or other values of α and β.

5.9 Examples.

Consider again the hypothesis testing problem of §4.9. Using the same notation as before, let $X_1, X_2, \ldots, X_{Kn}$ be independent normal random variables with unknown mean μ and unit variance. We wish to test

$$H_1^-: \mu<0 \ \ vs \ \ H_0: \mu=0 \ \ vs \ \ H_1^+: \mu>0$$

with error rates

$$\Pr(\mathcal{A}_0 \cup \mathcal{A}_1^+ | \mu=-0.5) = 0.1$$

$$\Pr(\mathcal{A}_1^- \cup \mathcal{A}_1^+ | \mu=0) = 0.05$$

$$\Pr(\mathcal{A}_1^- \cup \mathcal{A}_0 | \mu=0.5) = 0.1.$$

A maximum of 5 groups of 10 pairs of patients are available for entry on to the trial.

As a first example we consider the optimal wedge test for objective function $F_1: E(N|\mu=0)$. The standardized critical values for this test, $c_i' = c_i/\sqrt{in}$ and $l_i' = l_i/\sqrt{in}$, are given by: $l_1' = 0.0$, $c_1' = 4.001$, $l_2' = 0.665$, $c_2' = 3.122$, $l_3' = 1.03$, $c_3' = 2.720$, $l_4' = 1.351$, $c_4' = 2.404$ and $c_5' = 1.848$. It is interesting that the optimal wedge test does not permit the acceptance of H_0 at the first analysis. Under $\mu=0$ the expected sample size is 28.4, which compares very favourably with the relevant fixed sample size for this problem of 42.03 pairs of patients, and the optimal two-sided test for which $E(N|\mu=0) = 48.7$.

The optimal wedge test for $F_2: E(N|\mu=0.5)$ has standardized critical values: $l_1' = 0.0$, $c_1' = 2.613$, $l_2' = 0.266$, $c_2' = 2.920$, $l_3' = 0.838$, $c_3' = 2.385$, $l_4' = 1.425$,

$c_4' = 2.404$ and $c_5' = 2.243$. As with the first example it is optimal **not** to accept H_0 at the first analysis. As one would expect, both l_i' and c_i' are smaller here than in the first example when $i < 4$. Under $\mu = \pm 0.5$ the expected sample size of the test is 28.13 which is a slight improvement on the optimal two-sided test for this problem where $E(N|\mu=0.5) = 28.7$.

The optimal wedge test for objective function $F_5 : \int E(N|\mu)(0.5)^{-1}\varphi(\mu/0.5)d\mu$ has standardized critical values: $l_1' = 0.0$, $c_1' = 2.691$, $l_2' = 0.508$, $c_2' = 3.016$, $l_3' = 0.977$, $c_3' = 2.463$, $l_4' = 1.413$, $c_4' = 2.388$ and $c_5' = 2.070$. Here $F_5 = 33.3$ which compares favourably with the optimal two-sided test where $F_5 = 35.3$. As would be expected the stopping rule for this wedge test lies somewhere between the stopping rules for the first two examples.

It is interesting to note that the optimal wedge test for objective function $F_3 : E(N|\mu=2\delta)$ is identical to the optimal two-sided test given in §4.10. This is a result of the fact that under $\mu = 2\delta$, the gains in efficiency from being able to stop early and accept H_0 are small relative to the gains from accepting H_1^+.

5.10 A Comparison of Group Sequential Wedge Tests.

In this section we compare our optimal wedge tests with some of the wedge tests introduced in the literature and described in §5.3, and some of the two-sided tests of §4.

To simplify notation we let T_1^* denote the optimal wedge test for objective function $F_1 : E(N|\mu=0)$ over t with K fixed. Further we let T_1^+ denote the optimal two-sided test for F_1 with $t = 1.01$ and K fixed.

Table 5.4 gives F_1 expressed as a percentage of the fixed sample size, N_f, and, in parentheses, t, the ratio of the maximum sample size to N_f, for the tests of Gould & Pecore (1982) with $\pi_1 =$ 1.0, 0.5 and 0.15, and the optimal tests, T_1^* and T_1^+. Results are given for $\alpha = 0.05$, $\beta = 0.1$ and K = 2, 3, 4, 5 and 10, and, as before, they are independent of σ^2 and δ. The fixed sample size test for this problem has $K = 1$, $t = 1.0$ and $E(N|\mu=0) = \sigma^2 \text{x} 10.6/\delta^2$.

Table 5.4. $F_1 : E(N|\mu = 0)$ expressed as a percentage of fixed sample size, N_f, and, in parentheses, t, the maximum sample size expressed as a proportion of the fixed, for the tests of Gould & Pecore (1982) with π_1 = 1.0, 0.5 and 0.15, the optimal wedge test for F_1, T_1^*, and the optimal two-sided test for F_1 with $t = 1.01$, T_1^+, for $\alpha = 0.05$, $\beta = 0.1$ and K = 2, 3, 4, 5 and 10.

K	π_1 = 1.0	π_1 = 0.5	π_1 = 0.15	T_1^*	T_1^+
2	108.3	84.2	81.3	75.4	100.7
	(1.10)	(1.15)	(1.46)	(1.16)	(1.01)
3	112.8	78.8	81.4	71.1	100.6
	(1.15)	(1.35)	(2.13)	(1.34)	(1.01)
4	115.6	78.2	82.3	69.8	100.6
	(1.18)	(1.61)	(2.83)	(1.52)	(1.01)
5	117.7	80.0	82.7	67.8	100.5
	(1.21)	(1.91)	(3.54)	(1.21)	(1.01)
10	123.4	94.4	83.2	64.0	100.5
	(1.27)	(3.71)	(7.08)	(1.37)	(1.01)

By design the optimal wedge tests give the lowest value of F_1 for each K. The largest maximum sample size for an optimal wedge test occurs when $K = 4$ and equals 152% of the corresponding fixed sample size.

The Gould & Pecore test with π_1 = 1.0 is, as was mentioned in §5.3, identical to the two-sided test of Pocock (1977) (described in §4.4), and so does not permit the early acceptance of H_0. As one would expect, these tests have the largest values of F_1 and, with the exception of T_1^+, the lowest maximum sample sizes.

The other test without an inner wedge, T_1^+, also leads to large values of $E(N|\mu=0)$. For the examples given in the table it is always worse than the fixed sample size test. If we compare T_1^+ with the corresponding optimal wedge tests, the advantage of allowing the early acceptance of H_0 becomes clear.

Of the other two Gould & Pecore tests the one with $\pi_1 = 0.15$ is closest to optimal. It is never more than 20% of N_f from the overall minimum.

Both tests require very large maximum sample sizes (especially when $K = 10$) to meet the power requirements at $\mu=\pm\delta$. For example, the test with $\pi_1 = 0.15$ and $K = 10$ has a maximum sample size more than 7 times greater than N_f. In fact with π_1 reasonably small and $K > 3$ large maximum sample sizes would arguably render the Gould & Pecore tests impractical.

We have also compared wedge tests in terms of the objective function F_2: $E(N|\mu=\delta)$. Again to simplify notation we let T_2^* denote the optimal wedge test for F_2 over t with K fixed. Further we let T_2^+ denote the optimal two-sided test for F_2 over t with K fixed. Table 5.5 gives F_2 expressed as a percentage of the fixed sample size, N_f, and, in parentheses, t, for the tests of Gould & Pecore (1982) with $\pi_1 = 1.0$, 0.5, and 0.15, T_2^* and T_2^+. Again results are given for $\alpha = 0.05$, $\beta = 0.1$ and $K = 2$, 3, 4, 5 and 10, and they are independent of σ^2 and δ.

Table 5.5. $F_2: E(N|\mu=\delta)$ expressed as a percentage of fixed sample size, N_f, and, in parentheses, t, the maximum sample size expressed as a proportion of the fixed, for the tests of Gould & Pecore (1982) with π_1 = 1.0, 0.5 and 0.15, the optimal wedge test for F_2, T_2^*, and the optimal two-sided test for F_2, T_2^+, for $\alpha=0.05$, $\beta=0.1$ and K = 2, 3, 4, 5 and 10.

K	$\pi_1 = 1.0$	$\pi_1 = 0.5$	$\pi_1 = 0.15$	T_2^*	T_2^+
2	77.6 (1.10)	77.3 (1.15)	84.3 (1.46)	77.2 (1.12)	77.6 (1.10)
3	72.1 (1.15)	72.1 (1.35)	83.9 (2.13)	71.1 (1.22)	72.1 (1.14)
4	69.7 (1.18)	70.4 (1.61)	84.2 (2.83)	68.1 (1.24)	69.6 (1.15)
5	68.5 (1.21)	70.1 (1.91)	84.3 (3.54)	66.2 (1.30)	68.1 (1.16)
10	66.6 (1.27)	72.0 (3.71)	84.5 (7.08)	62.4 (1.44)	65.1 (1.17)

Again, by design, the wedge tests give the lowest values of F_2. Both the Pocock test (π_1 = 1.0) and the optimal two-sided test are close to the optimal wedge test when K is small. They also require somewhat smaller maximum sample sizes than T_2^*. However with K = 10 the wedge test is substantially better than the other five designs.

Of the other two Gould & Pecore tests the one with π_1 = 0.5 is clearly the best in terms of giving low values of F_2. As π_1 decreases the critical values of the inner wedge, l_i, increase which, in turn, forces the outer critical values, c_i, to increase. This leads to a smaller probability of stopping early under $\mu = \delta$ and

explains the inefficiency of the tests with $\pi_1 = 0.15$.

The maximum sample sizes for the Gould & Pecore tests are, of course, the same as in Table 5.4. Hence the earlier criticisms remain concerning the need for large maximum sample sizes when π_1 is small and K is large.

The tests of Emerson & Fleming (1989) cannot be compared directly with the tests of Tables 5.4 and 5.5 as, in general, they do not satisfy the Type II error constraints exactly. We can, however, compare any given Emerson & Fleming test with an optimal test with the same errors.

To ease notation, let r_μ be the ratio of $E(N|\mu)$ for the optimal test to $E(N|\mu)$ for the Emerson & Fleming test, expressed as a percentage. Clearly $r_\mu \leq 100\%$, with $r_\mu = 100\%$ if the Emerson & Fleming test is optimal. Table 5.6 gives r_0 and r_δ for the Emerson & Fleming tests with $p = 0$ and $p = 0.5$, $K = 2$, 3, 4, 5 and 10 and $\alpha = 0.05$. Also given, in parentheses, are the attained Type II errors under $\mu = \delta$. The results given are independent of σ^2 and δ.

Table 5.6. The efficiency ratios r_0 and r_δ, and, in parentheses, the Type II errors under $\mu = \delta$, for the tests of Emerson & Fleming (1989) with $p = 0$ and $p = 0.5$, K = 2, 3, 4, 5 and 10 and $\alpha = 0.05$.

	p = 0		p = 0.5	
K	r_0	r_δ	r_0	r_δ
2	77.5 (0.024)	87.8 (0.024)	97.5 (0.025)	95.6 (0.025)
3	92.5 (0.025)	82.7 (0.025)	95.7 (0.024)	94.8 (0.024)
4	86.8 (0.024)	81.7 (0.024)	89.3 (0.018)	90.2 (0.018)
5	89.9 (0.025)	81.7 (0.025)	91.5 (0.019)	91.2 (0.019)
10	88.3 (0.025)	80.4 (0.025)	90.4 (0.017)	90.2 (0.017)

The Emerson & Fleming tests with $p = 0$ are between 77.5% and 92.5% efficient under $\mu = 0$, and between 80.4% and 87.8% efficient under $\mu = \delta$. The observed Type II errors range from 0.0237 up to 0.0249.

The tests with $p = 0.5$ can be seen to be uniformly closer to optimal than those with $p = 0$. Under $\mu = \delta$ the tests are between 90.2% and 95.6% efficient, while under $\mu = 0$ efficiency ratios are particularly impressive, ranging from 89.3% up to 97.5%. Observed Type II errors are between 0.017 and 0.025.

For any given problem, with K and α fixed, we can minimize a chosen objective function over the family of Emerson & Fleming tests, by searching over the parameter p. This has been done for objective functions $F_1: E(N|\mu{=}0)$ and $F_2: E(N|\mu{=}\delta)$, $\alpha = 0.05$ and K = 2, 3, 4, 5 and 10. Table 5.7 gives r_0 and r_δ for

the "optimal" Emerson & Fleming test for each problem. Also given, in parentheses, are the attained Type II errors under $\mu = \delta$.

Table 5.7. The efficiency ratios r_0 and r_δ, and, in parentheses, the Type II errors under $\mu = \delta$, for the "optimal" tests of Emerson & Fleming (1989) for K = 2, 3, 4, 5 and 10 and $\alpha = 0.05$.

K	r_0	r_δ
2	97.7 (0.025)	97.0 (0.025)
3	98.7 (0.025)	94.8 (0.024)
4	96.1 (0.024)	91.5 (0.02)
5	95.9 (0.022)	92.1 (0.02)
10	95.0 (0.023)	90.6 (0.018)

The results are rather impressive, with not one of the "optimal" Emerson & Fleming tests being less than 90% efficient. We note however that the computer time necessary to compute an "optimal" Emerson & Fleming test is not substantially less than that required to compute an overall optimal test using our improved method.

5.11 The Bioequivalence Problem.

As has already been mentioned, one of the most important applications of wedge tests is in the area of bioequivalence tests. in this section we describe the bioequivalence problem and the reasons why optimal wedge tests are particularly suitable here.

Suppose that all patients suffering from a particular disease are given a standard drug, S. A new drug, N, has been developed which it is thought will result in patients experiencing fewer side-effects than is the case with S. The two drugs are chemical equivalents; that is, as defined by Metzler (1974), they are drugs '... of the same dosage form which contain equal amounts of the same active ingredient as indicated by official standards'. Chemical equivalence on its own does not guarantee that the two drugs will be equally effective in the treatment of patients. To try to establish whether S and N are equally effective we could conduct a clinical trial similar to those described in §4.2 and §5.2. However such trials are complicated and expensive to run. Metzler (1974) suggested that logistical considerations often lead us to investigate the biological equivalence of S and N. He stated that biological equivalents are '... those chemical equivalents which deliver the same amount of active ingredient into the circulating blood'. Drugs which are biological equivalents are assumed to be equally effective.

There are a number of possible patient responses one might record when testing for bioequivalence. Racine-Poon *et al.* (1986) and (1987) suggested that one might record the logarithm of the ratio of the areas under the blood level concentration curves for S and N, which they denoted by Ψ. Most of the literature on bioequivalence assumes that a two-period crossover design is employed by the experimenter. Here patients are randomly assigned to one of the two treatments on entry to the trial. Their blood level concentration is then monitored over time. A wash-out period follows to remove the effects of the drug, and then the patients are given the treatment they did not receive in the first period.

Racine-Poon *et al.* (1986) and (1987) considered a Bayesian approach to the bioequivalence problem. The parameters Ψ and σ^2 (the variance of the patient responses which is assumed unknown) are assigned a vague prior distribution.

After the experiment has finished the posterior distribution for Ψ is formed and the bioequivalence of the two drugs accepted if

$$\Pr(0.8 \leq \Psi \leq 1.2 \mid y) \geq 0.95$$

where y represents the data.

Racine-Poon *et al.* noted that typical bioequivalence trials are conducted with as few as 6 patients. For most vague priors such a small sample is unlikely to lead to the acceptance of bioequivalence even when $\Psi = 1$. To deal with this problem they suggested a two-stage sequential approach. In the first stage n_1 patients (where n_1 is typically around 6) are admitted on to the trial. On the basis of their responses the probability of accepting bioequivalence is predicted. If this probability is suitably high (low) the trial is stopped and bioequivalence is accepted (rejected). Otherwise the second group of n_2 patients is entered on to the trial and after they have responded the trial is terminated with the acceptance or rejection of bioequivalence.

This two-stage procedure has a markedly higher probability of accepting bioequivalence when Ψ is close to one than does the single sample approach. We note that it is also a Bayesian "inner wedge" design in so far as it allows the acceptance or rejection of bioequivalence at the first stage.

We now consider a frequentist two-stage approach to bioequivalence. Suppose $Y_1, Y_2, \ldots$ are independent normal random variables with mean Ψ and known variance τ^2 and we wish to test

$$H_1^-: \Psi < 1 \quad vs \quad H_0: \Psi = 1 \quad vs \quad H_1^+: \Psi > 1$$

with error rates

$$\Pr(\mathcal{A}_0 \cup \mathcal{A}_1^+ \mid \Psi = 0.8) = \beta$$

$$\Pr(\mathcal{A}_1^- \cup \mathcal{A}_1^+ \mid \Psi = 1) = \alpha$$

$$\Pr(\mathcal{A}_1^- \cup \mathcal{A}_0 \mid \Psi = 1.2) = \beta.$$

To tie this problem in with the symmetric one posed in §5.2 we consider the following transformation:

Let

$$X_i = (Y_i - 1)\frac{\delta}{0.2} \qquad (i = 1, 2, \ldots)$$

then $X_i \sim N(\mu,\sigma^2)$, where

$$\mu = (\Psi-1)\frac{\delta}{0.2} \quad \text{and} \quad \sigma^2 = \frac{\delta^2\,\tau^2}{0.04}.$$

and we wish to test

$$H_1^-: \mu < 0 \quad vs \quad H_0: \mu = 0 \quad vs \quad H_1^+: \mu > 0$$

with error rates given by equations (5.2.1)-(5.2.3). Optimal stopping rules for this problem are easily transformed for use with the first problem (see Appendix 5.2 for details).

Before continuing it is worth pausing to comment on some unusual features of this hypothesis testing problem. Jennison (1986) in the discussion of Racine-Poon *et al.* (1986) questioned the use of a two-sided test for this problem. He felt that as we are testing a new formulation against a standard, a one-sided test would be more appropriate. Racine-Poon *et al.* countered by pointing out that if the null hypothesis is false then either too little of the active ingredient of the drug is released into the blood stream and the new drug may be relatively ineffective, or too much of the active ingredient is released which may produce a toxic effect. Rejection of H_0 in either direction should clearly lead to the termination of the development of the new drug.

In all the previous examples we have considered, acceptance of H_0 has resulted in a passive action - the continuing use of the standard treatment. In the bioequivalence problem, however, the roles of the null and alternative hypotheses have swapped. Now acceptance of H_0 would lead to the possibility of a change in treatment regimens with the new treatment replacing the standard. The roles of our Type I and Type II errors have also swapped. To reject the null hypothesis when it is true (a Type I error) would cost the drug company in needlessly lost profits and effect patients in the sense that they do not receive the new drug with the less harmful side-effects. To accept H_0 when it is false (a Type II error) is clearly unethical. This change in roles of the error rates has to be borne in mind when designing trials.

Bioequivalence testing is an area where a Bayesian approach is attractive because of the large amount of prior information which is likely to be available to the experimenter. Jennison (1986) has suggested that this prior information might be used in the design of optimal frequentist tests. His idea was to produce an

optimal test which minimizes expected sample size integrated over the experimenter's prior. The optimal tests for F_5 are examples of such designs.

We now proceed to consider the optimal frequentist designs for the following priors:

$$(i)\ \pi_1(\mu) \sim N(0,4\delta^2)$$

$$(ii)\ \pi_2(\mu) \sim N(0,9\delta^2).$$

We shall denote the optimal tests which minimize expected sample size integrated over $\pi_1(\mu)$ and $\pi_2(\mu)$ by F_6 and F_7 respectively. We note that the optimal tests for F_6 correspond to the case where the experimenter's *a priori* beliefs are vaguer than was the case with F_5. Objective function F_7 corresponds to a still vaguer set of prior beliefs.

Table 5.8 gives the minima of F_6 and F_7 for $K = 2$, $t = 1.01$, 1.05, 1.1 and 1.15, $\alpha = 0.05$ and 0.1 and $\beta = 0.01$, 0.025 and 0.05. The results given are independent of δ and σ^2.

Table 5.8. Minima of F_6 and F_7 expressed as percentages of the fixed sample size, N_f, for $K = 2$, $t = 1.01$, 1.05, 1.1 and 1.15, $\alpha = 0.05$ and 0.1, and $\beta = 0.01$, 0.025 and 0.05.

		F_6				F_7			
		t				t			
		1.01	1.05	1.1	1.15	1.01	1.05	1.1	1.15
$\alpha = 0.05$	$\beta = 0.01$	67.4	66.2	66.5	67.6	62.1	61.8	62.8	64.4
	$\beta = 0.025$	68.9	67.1	67.2	68.2	63.2	62.5	63.3	64.8
	$\beta = 0.05$	70.5	68.0	67.9	68.7	64.3	63.2	63.9	65.2
$\alpha = 0.1$	$\beta = 0.01$	67.0	66.5	67.1	68.2	61.8	62.0	63.2	64.7
	$\beta = 0.025$	68.6	67.7	67.9	68.9	62.9	62.9	63.8	65.2
	$\beta = 0.05$	70.2	68.9	68.8	69.6	64.1	63.8	64.5	65.8

Even though we have confined ourselves to two-stage procedures, Table 5.8 shows that very efficient designs may be obtained. About 33% of the fixed sample size is saved in each problem. Even when t and β are very small the gains in efficiency are large.

Our results show that it is possible to obtain an optimal frequentist group sequential design for the bioequivalence problem which makes use of the experimenter's *a priori* beliefs when choosing a suitable objective function. The opportunity to stop sampling after the first stage with the acceptance or rejection of bioequivalence make our designs particularly attractive.

5.12 Discussion and Conclusions.

In §5 we have considered the generalization of the two-sided group sequential designs of §4 to allow for the early acceptance of H_0. Results show that such a generalization can lead to large gains in efficiency particularly when looking at optimal tests for objective function F_1: $E(N|\mu=0)$.

An important application of wedge designs is in the area of bioequivalence testing. In §5.11 we considered optimal wedge designs for this problem and showed that, even when logistical considerations limit us to two-stage procedures with low maximum sample sizes, very efficient designs may be computed.

Appendix.

Appendix 5.1.

This Appendix is similar to Appendix 3.1 where we gave details of how to calculate numerically the power function and the expected sample size function for a one-sided group sequential test. Here we give similar details for wedge tests. As a two-sided group sequential test is just a special case of a wedge test, the work of this Appendix applies to the tests of §4 as well.

As an example consider calculating the operating characteristic function, $OC(\mu) = \Pr(\mathcal{A}_0 | \mu)$, given by equation (5.A1.1)

$$OC(\mu) = \sum_{j=1}^{K} \int_{-l_j}^{l_j} \int_{r_{j-1}} \cdots \int_{r_1} f_\mu(x_1)\, f_\mu(x_2 - x_1) \cdots f_\mu(x_j - x_{j-1})\, dx_1 \ldots dx_{j-1}\, dx_j \qquad (5.A1.1)$$

where $r_i = \{(-c_i, -l_i) \cup (l_i, c_i)\}$ for $i = 1, 2, \ldots, K-1$, and $f_\mu(x)$ is a normal density with mean $n\mu$ and variance $n\sigma^2$.

Clearly $OC(\mu)$ is the sum of an integral and $K-1$ multiple integrals. The first term in the sum is simply

$$\int_{-l_1}^{l_1} f_\mu(x_1)\, dx_1$$

which equals

$$\Phi\left[\frac{l_1 - n\mu}{\sqrt{n\sigma^2}}\right] - \Phi\left[\frac{-l_1 - n\mu}{\sqrt{n\sigma^2}}\right]$$

where Φ is the c.d.f. of the standard normal distribution. The NAG library contains a subroutine, S15ABF, for calculating $\Phi(.)$.

The second term in the sum is

$$\int_{-l_2}^{l_2} \int_{r_1} f_\mu(x_1)\, f_\mu(x_2 - x_1)\, dx_1\, dx_2 \qquad (5.A1.2)$$

The integral with respect to x_1 here can be evaluated using Simpson's rule. One approach would be to use a fixed number of equally spaced grid points. However, as was pointed out in Appendix 3.1, for a multi-use algorithm such an approach might lead to inaccurate calculations.

An improved approach involves placing a grid of $6N-1$ points $\{g_i: 1 \le i \le 6N-1\}$, where N is fixed, over x_1 according to the rule:
For $i = 1, N$

$$g_i = n\mu - \sqrt{n}\,\sigma\{3+4\ln(N/i)\}$$

For $i = N+1, 5N-1$

$$g_i = n\mu - \sqrt{n}\,\sigma\{3-6(i-N)/4N\}$$

and, for $i = 5N, 6N-1$,

$$g_i = n\mu + \sqrt{n}\,\sigma\{3+4\ln(i/(6N-1))\}.$$

This rule places a fixed number ($4N$) of equally spaced points within 3 standard deviations of the mean of the distribution we are integrating. In the tails of the distribution grid points are placed increasingly far apart. To obtain a set of grid points, $\{g_{1,i1}: 1 \le i1 \le N_1\}$ to place over the interval r_1 we first find g_i and g_{i+1} such that

$$g_i \le -c_1 < g_{i+1},$$

we next find $g_{i'}$ and $g_{i'+1}$ such that

$$g_{i'} < -l_1 \le g_{i'+1}$$

and then g_j and g_{j+1} such that

$$g_j \le l_1 < g_{j+1}.$$

and finally $g_{j'}$ and $g_{j'+1}$ such that

$$g_{j'} < c_1 \le g_{j'+1}.$$

We then set $g_{1,1} = -c_1$, $g_{1,3} = g_{i+1}$, $g_{1,5} = g_{i+2}$, ..., $g_{1,M1-2} = g_{i'-1}$, $g_{1,M1} = -l_1$, $g_{1,M1+1} = l_1$, $g_{1,M1+3} = g_{j+1}$, ..., $g_{1,N_1-2} = g_{j'}$ and $g_{1,N_1} = c_1$. (Note that if $-c_1 < g_1$ and/or $c_1 > g_{6N-1}$, we set $g_{1,1} = g_1$ and/or $g_{1,N1} = g_{6N-1}$.) The grid points with even subscripts $\{g_{1,2i}: 1 \le i \le (N_1-1)/2\}$ are then positioned halfway between the neighbouring grid points with odd subscripts, i.e.

$$g_{1,2i} = \frac{1}{2}\{g_{1,2i-1} + g_{1,2i+1}\}.$$

The weights $\{w_{1,i1}: 1 \le i1 \le N_1\}$ for use in integral (5.A1.2) are given by

$$w_{1,1}=\frac{1}{6}(g_{1,3}-g_{1,1})$$

$$w_{1,2i} = \frac{4}{6}(g_{1,2i+1}-g_{1,2i-1}) \quad i=1,2,\ldots,(M1-1)/2$$

$$w_{1,2i+1} = \frac{1}{6}(g_{1,2i+3}-g_{1,2i-1}) \quad i=1,2,\ldots,(M1-3)/2$$

$$w_{1,M1} = \frac{1}{6}(g_{1,M1}-g_{1,M1-2})$$

$$w_{1,M1+1} = \frac{1}{6}(g_{1,M1+3}-g_{1,M1+2})$$

$$w_{1,2i+1} = \frac{4}{6}(g_{1,2i+2}-g_{1,2i}) \quad i=(M1+1)/2,\ldots,(N1-2)/2$$

$$w_{1,2i} = \frac{1}{6}(g_{1,2i+2}-g_{1,2i-2}) \quad i=(M1+3)/2\ldots,(N1-2)/2$$

$$w_{1,N1} = \frac{1}{6}(g_{1,M1}-g_{1,M1-2}).$$

Then the multiple integral (5.A1.2) is approximately equal to

$$\int_{-l_2}^{l_2} \sum_{i1=1}^{N1} w_{1,i1}\, f_\mu(g_{1,i1})\, f_\mu(x_2-g_{1,i1})\, dx_2 \qquad (5.A1.3)$$

at this stage in the computations we store the terms $\{w_{1,i1} f_\mu(g_{1,i1}): 1 \le i1 \le N1\}$ in an array $\{h_1(g_{1,i1}): 1 \le i1 \le N1\}$ for future use.
The integral (5.A1.3) is equal to

$$\sum_{i1=1}^{N1} h_1(g_{1,i1})\left\{\Phi\left[\frac{l_2-g_{1,i1}-n\mu}{\sqrt{n}\,\sigma}\right] - \Phi\left[\frac{-l_2-g_{1,i1}-n\mu}{\sqrt{n}\,\sigma}\right]\right\}.$$

This sum is easily evaluated using the stored array $\{h_1(g_{1,i1}): 1 \le i1 \le N1\}$ and the NAG library subroutine S15ABF.

The third term in the sum (5.A1.1) is the multiple integral

$$\int_{-l_3}^{l_3} \int_{r_2} \int_{r_1} f_\mu(x_1)\, f_\mu(x_2-x_1)\, f_\mu(x_3-x_2)\, dx_1\, dx_2\, dx_3. \qquad (5.A1.4)$$

Clearly the first stage in evaluating (5.A1.4) is to evaluate the integral with respect to x_1. This is easily achieved by using the stored array $\{h_1(g_{1,i1}): 1 \le i1 \le N1\}$.

So the multiple integral (5.A1.4) is approximately equal to

$$\int_{-l_3}^{l_3}\int_{r_2}\sum_{i1=1}^{N1} h_1(g_{1,i1})\, f_\mu(x_2-g_{1,i1})\, f_\mu(x_3-x_2)\, dx_2\, dx_3. \qquad (5.A1.5)$$

We can now use Simpson's rule to evaluate the integral with respect to x_2. The relevant grid points and weights are obtained by using the same rule as for the integral with respect to x_1. We shall denote the grid points by $\{g_{2,i2}: 1 \le i2 \le N2\}$ and the weights by $\{w_{2,i2}: 1 \le i2 \le N2\}$. Integral (5.A1.5) is then approximately equal to

$$\int_{-l_3}^{l_3}\sum_{i2=1}^{N2}\sum_{i1=1}^{N1} h_1(g_{1,i1})\, w_{2,i2}\, f_\mu(g_{2,i2}-g_{1,i1})\, f_\mu(x_3-g_{2,i2})\, dx_3. \qquad (5.A1.6)$$

We store the terms $\{\sum_{i1=1}^{N1} h_1(g_{1,i1})\, w_{2,i2}\, f_\mu(g_{2,i2}-g_{1,i1}): 1 \le i2 \le N2\}$ in an array $\{h_2(g_{2,i2}): 1 \le i_2 \le N2\}$ for future use. It follows that (5.A1.6) is equal to

$$\sum_{i2=1}^{N2} h_2(g_{2,i2})\left\{\Phi\left[\frac{l_3-g_{2,i2}-n\mu}{\sqrt{n}\sigma}\right] - \Phi\left[\frac{-l_3-g_{2,i2}-n\mu}{\sqrt{n}\sigma}\right]\right\}$$

This sum is easily evaluated using the stored array $\{h_2(g_{2,i2}): 1 \le i2 \le N2\}$ and the NAG library subroutine S15ABF.

All other terms in the sum (5.A1.1) are calculated similarly. The Kth term equals

$$\sum_{i(K-1)=1}^{N(K-1)} h_{K-1}(g_{K-1,i(K-1)})\left\{\Phi\left[\frac{l_K-g_{K-1,i(K-1)}-n\mu}{\sqrt{n}\sigma}\right] - \Phi\left[\frac{-l_K-g_{K-1,i(K-1)}-n\mu}{\sqrt{n}\sigma}\right]\right\}$$

where $\{h_{K-1}(g_{K-1,i(K-1)}): 1 \le i(K-1) \le N(K-1)\}$ is a stored array with $i(K-1)$st element

$$h_{K-1}(g_{K-1,i(K-1)}) = \sum_{i(K-2)=1}^{N(K-2)} h_{K-2}(g_{K-2,i(K-2)})\, w_{K-1,i(K-1)}\, f_\mu(g_{K-1,i(K-1)}-g_{K-2,i(K-2)}).$$

Using the same techniques as those explained above we can calculate the expected sample size function and the power function of a given group sequential wedge test. Details are omitted.

Calculations for the expected sample size function, power function and operating characteristic function of a given two-sided group sequential test are similar to those described here, with $l_i = 0$ of course.

Appendix 5.2.

Consider again the group sequential inner wedge problem introduced in §5.2, which has a maximum of K groups of n pairs of patients. Using the same notation as before, let $X_1, X_2, \ldots, X_{Kn}$ be independent normal random variables with unknown mean μ and known variance σ^2. We wish to test

$$H_1^-: \mu < 0 \quad vs \quad H_0: \mu = 0 \quad vs \quad H_1^+: \mu > 0$$

with error rates given by equations (5.2.1)-(5.2.3).

Suppose the set of critical values $\{(l_1,c_1),(l_2,c_2),\ldots,(l_{K-1},c_{K-1}),c_K\}$ defines a feasible stopping rule for the above problem. The expected sample size function for this test is given by equation (5.2.5). Further, the operating characteristic function for this problem, $OC(\mu) = \Pr(\mathcal{A}_0|\mu)$, is given by equation (5.A2.1)

$$OC(\mu) = \sum_{j=1}^{K} \int_{-l_j}^{l_j} \int_{r_{j-1}} \cdots \int_{r_1} f_\mu(x_1)\, f_\mu(x_2-x_1) \cdots f_\mu(x_j-x_{j-1})\, dx_1 \ldots dx_{j-1}\, dx_j \quad (5.A2.1)$$

where $r_i = \{(-c_j,-l_j)\cup(l_j,c_j)\}$ for $i = 1, 2, \ldots, K-1$, $l_K = c_K$, and $f_\mu(x)$ is a normal density with mean $n\mu$ and variance $n\sigma^2$.

Finally the power function, $\pi_1(\mu) = \Pr(\mathcal{A}_1^+|\mu)$, is given by equation (5.A2.2)

$$\pi_1(\mu) = \sum_{j=1}^{K} \int_{c_j}^{\infty} \int_{r_{j-1}} \cdots \int_{r_1} f_\mu(x_1)\, f_\mu(x_2-x_1) \cdots f_\mu(x_j-x_{j-1})\, dx_1 \ldots dx_{j-1}\, dx_j \quad (5.A2.2)$$

We now consider a new problem in which the variance of the original problem, σ^2, is replaced by ${\sigma_1}^2$, and δ is replaced by δ_1 (>0). The maximum number of analyses for this new test, K, and α and β remain the same as for the original problem.

Consider the following stopping rule for our new problem defined by the rescaled set of critical values $\{({l_1}^*,{c_1}^*), ({l_2}^*,{c_2}^*), \ldots, ({l_{K-1}}^*,{c_{K-1}}^*), {c_K}^*\}$, where

$l_i^* = \frac{\sigma_1^2 \delta}{\sigma^2 \delta_1} l_i$ $(i = 1, 2, \ldots, K-1)$ and $c_i^* = \frac{\sigma_1^2 \delta}{\sigma^2 \delta_1} c_i$ $(i = 1, 2, \ldots, K)$, together with the rescaled group sizes $n^* = \frac{\sigma_1^2 \delta^2}{\sigma^2 \delta_1^2} n$.

In this Appendix we prove the following Lemmas and Corollaries:

Lemma 5.1 : The operating characteristic function for our new problem, $OC^*(\mu)$, is such that

$$OC^*(\mu) = OC(\mu\delta/\delta_1).$$

Corollary 5.1 : From **Lemma 5.1** it follows that

$$OC^*(0) = OC(0) = 1-\alpha.$$

Lemma 5.2 : The power function for our new problem, $\pi_1^*(\mu)$, is such that

$$\pi_1^*(\mu) = \pi_1(\mu\delta/\delta_1).$$

Corollary 5.2 : From **Lemma 5.2** it follows that

$$\pi_1^*(\delta_1) = \pi_1(\delta)$$

and

$$\pi_1^*(-\delta_1) = \pi_1(-\delta).$$

As $\pi_1(\delta) = 1-\beta$ and $\pi_1(-\delta) = \beta$, it follows from Corollaries 5.1 and 5.2 that the rescaled set of critical values $\{(l_1^*, c_1^*), \ldots, (l_{K-1}^*, c_{K-1}^*), c_K^*\}$ together with the rescaled group sizes, n^*, define a feasible stopping rule for our new problem.

Lemma 5.3 : Letting $E(N^*|\mu)$ denote the expected sample size function for our new problem and N_f^* denote the corresponding fixed sample size, we prove that

$$\frac{E(N^*|\mu)}{N_f^*} = \frac{E(N|\mu\delta/\delta_1)}{N_f}$$

where $E(N|\mu)$ is the expected sample size function for the original problem and

N_f is the corresponding fixed sample size.

Corollary 5.3 : From **Lemma 5.3** it follows that

$$(i)\quad \frac{E(N^*|\mu=0)}{N_f^*} = \frac{E(N|\mu=0)}{N_f}$$

$$(ii)\quad \frac{E(N^*|\mu=\delta_1)}{N_f^*} = \frac{E(N|\mu=\delta)}{N_f}$$

$$(iii)\ \frac{\int E(N^*|\mu)\,\delta_1^{-1}\,\varphi(\mu/\delta_1)d\mu}{N_f^*} = \frac{\int E(N|\mu')\delta^{-1}\,\varphi(\mu'/\delta)d\mu'}{N_f}$$

where $\mu' = \mu\delta/\delta_1$.

Proof of Lemma 5.1 :

The operating characteristic function for the new problem is given by

$OC^*(\mu) =$

$$\sum_{j=1}^{K} \int_{-l_j^*}^{l_j^*} \int_{r_{j-1}^*} \cdots \int_{r_1^*} f_\mu(x_1)\, f_\mu(x_2-x_1) \cdots f_\mu(x_j-x_{j-1})\, dx_1 \cdots dx_{j-1}\, dx_j \quad (5.A2.3)$$

where $r_i^* = \{(-c_i^*,-l_i^*)\cup(l_i^*,c_i^*)\}$ for $i=1,2,\ldots,K-1$, and $f_\mu(x)$ is a normal density with mean $n^*\mu$ and variance $n^*{\sigma_1}^2$.

Consider the substitutions $z_i = \dfrac{\sigma^2\delta_1}{{\sigma_1}^2\delta} x_i$ for $i=1,2,\ldots,K$, in equation (5.A2.3). We obtain

$OC^*(\mu) =$

$$\sum_{j=1}^{K} \int_{-l_j}^{l_j} \int_{r_{j-1}} \cdots \int_{r_1} g_{\mu'}(z_1)\, g_{\mu'}(z_2-z_1) \cdots g_{\mu'}(z_j-z_{j-1})\, dz_1 \cdots dz_{j-1}\, dz_j \quad (5.A2.4)$$

where $r_i = \{(-c_i,-l_i)\cup(l_i,c_i)\}$ for $i=1,2,\ldots,K-1$, $\mu' = \mu\delta/\delta_1$, $g_{\mu'}(z)$ is a normal density with mean $n\mu'$ and variance $n\sigma^2$, and $\{(l_1,c_1),\ldots,(l_{K-1},c_{K-1}),c_K\}$ is a feasible set of critical values for our original problem.

Comparing the RHS of equation (5.A2.4) with equation (5.A2.1), we see that

$$OC^*(\mu) = OC(\mu') = OC(\mu\delta/\delta_1). \quad (5.A2.5)$$

Q.E.D.

Proof of Corollary 5.1 :

Substituting $\mu = 0$ in to equation (5.A2.5) gives

$$OC^*(0) = OC(0),$$

and from equation (5.2.1) we have $OC(0) = 1-\alpha$. Therefore $OC^*(0) = 1-\alpha$.

Proof of Lemma 5.2 :

The power function for the new problem is given by

$\pi_1{}^*(\mu) =$

$$\sum_{j=1}^{K} \int_{c_j{}^*}^{\infty} \int_{r_{j-1}{}^*} \cdots \int_{r_1{}^*} f_\mu(x_1)\, f_\mu(x_2-x_1) \cdots f_\mu(x_j-x_{j-1})\, dx_1 \cdots dx_{j-1}\, dx_j \quad (5.A2.6)$$

where, as in **Lemma 5.1**, $r_i{}^* = \{(-c_i{}^*,-l_i{}^*)\cup(l_i{}^*,c_i{}^*)\}$ for $i=1,2,\ldots,K-1$, and $f_\mu(x)$ is a normal density with mean $n^*\mu$ and variance $n^*\sigma_1{}^2$.

Again, we consider the substitutions $z_i = \dfrac{\sigma^2\delta_1}{\sigma_1{}^2\delta} x_i$ for $i=1,2,\ldots,K$, this time in equation (5.A2.6). We obtain

$\pi_1{}^*(\mu) =$

$$\sum_{j=1}^{K} \int_{c_j}^{\infty} \int_{r_{j-1}} \cdots \int_{r_1} g_{\mu'}(z_1)\, g_{\mu'}(z_2-z_1) \cdots g_{\mu'}(z_j-z_{j-1})\, dz_1 \cdots dz_{j-1}\, dz_j \quad (5.A2.7)$$

where $r_i = \{(-c_i,-l_i)\cup(l_i,c_i)\}$ for $i=1,2,\ldots,K-1$, $\mu' = \mu\delta/\delta_1$, $g_{\mu'}(z)$ is a normal density with mean $n\mu'$ and variance $n\sigma^2$ and $\{(l_1,c_1),\ldots,(l_{K-1},c_{K-1}),c_K\}$ is a feasible set of critical values for our original problem.

Comparing the RHS of equation (5.A2.7) with equation (5.A2.2), we see that

$$\pi_1{}^*(\mu) = \pi_1(\mu') = \pi_1(\mu\delta/\delta_1) \quad (5.A2.8)$$

Q.E.D.

Proof of Corollary 5.2 :

Substituting $\mu = \delta_1$ in to equation (5.A2.8) gives

$$\pi_1{}^*(\delta_1) = \pi_1(\delta),$$

and from equation (5.2.3) we have $\pi_1(\delta) = \beta$. Therefore $\pi_1{}^*(\delta_1) = \beta$.

Substituting $\mu = -\delta_1$ in to equation (5.A2.8) gives

$$\pi_1^*(-\delta_1) = \pi_1(-\delta),$$

and from equation (5.2.2) we have $\pi_1(-\delta) = 1-\beta$. Therefore $\pi_1^*(-\delta_1) = 1-\beta$.

Proof of Lemma 5.3 :

The expected sample size function for our new problem is given by

$E(N^*|\mu) =$

$$n^* \sum_{j=1}^{K} j \int_{s_j^*} \int_{r_{j-1}^*} \cdots \int_{r_1^*} f_\mu(x_1) f_\mu(x_2-x_1) \cdots f_\mu(x_j-x_{j-1})\, dx_1 \ldots dx_{j-1} dx_j \quad (5.A2.9)$$

where $s_j^* = \{(-\infty,-c_j^*) \cup (-l_j^*, l_j^*) \cup (c_j^*,\infty)\}$ for $j = 1, 2, \ldots, K$, $r_i^* = \{(-c_i^*,-l_i^*) \cup (l_i^*, c_i^*)\}$ for $i = 1, 2, \ldots, K-1$, and $f_\mu(x)$ is a normal density with mean $n^*\mu$ and variance $n^*\sigma_1^2$.

Consider the substitutions $z_i = \dfrac{\sigma^2 \delta_1}{\sigma_1^2 \delta} x_i$ for $i = 1, 2, \ldots, K$, in equation (5.A2.9). We obtain

$E(N^*|\mu) =$

$$n^* \sum_{j=1}^{K} j \int_{s_j} \int_{r_{j-1}} \cdots \int_{r_1} g_{\mu'}(z_1)\, g_{\mu'}(z_2-z_1) \cdots g_{\mu'}(z_j-z_{j-1})\, dz_1 \ldots dz_{j-1} dz_j \quad (5.A2.10)$$

where $s_j = \{(-\infty,-c_j) \cup (-l_j, l_j) \cup (c_j,\infty)\}$ for $j = 1, 2, \ldots, K$, $r_i = \{(-c_i,-l_i) \cup (l_i, c_i)\}$ for $i = 1, 2, \ldots, K-1$, $g_{\mu'}(x)$ is a normal density with mean $n\mu'$ and variance $n\sigma^2$, and $\{(l_1,c_1), \ldots, (l_{K-1},c_{K-1}), c_K\}$ is a feasible set of critical values for our original problem.

Comparing the RHS of equation (5.A2.10) with equation (5.2.5), we have

$$E(N^*|\mu) = \frac{n}{n^*} E(N|\mu') \quad (5.A2.11)$$

where $E(N|\mu')$ is the expected sample size function of our original problem under a treatment difference of μ'.

From earlier we have

$$n^* = \frac{\sigma_1^2 \delta^2}{\sigma^2 \delta_1^2} n.$$

Substituting for n^* in equation (5.A2.11) and rearranging, we obtain

$$\frac{E(N^*|\mu)}{\sigma_1{}^2\delta^2} = \frac{E(N|\mu')}{\sigma^2\delta_1^2} \qquad (5.A2.12)$$

and dividing both sides of equation (5.A2.12) by $\{\Phi^{-1}(1-\beta) + \Phi^{-1}(1-\alpha/2)\}^2/\delta^2\delta_1{}^2$ gives

$$\frac{E(N^*|\mu)}{N_f^*} = \frac{E(N|\mu')}{N_f} \qquad (5.A2.13)$$

where, by analogy with equation (5.2.4),

$$N_f^* = \frac{\sigma_1{}^2}{\delta_1{}^2}\{\Phi^{-1}(1-\beta) + \Phi^{-1}(1-\alpha/2)\}^2.$$

Q.E.D.

Proof of Corollary 5.3 :

(i) Substituting $\mu = 0$ in to equation (5.A2.13) gives

$$\frac{E(N^*|\mu=0)}{N_f^*} = \frac{E(N|\mu=0)}{N_f}$$

(ii) Substituting $\mu = \delta_1$ in to equation (5.A2.13) gives

$$\frac{E(N^*|\mu=\delta_1)}{N_f^*} = \frac{E(N|\mu=\delta)}{N_f}$$

(iii) Consider substituting $\mu' = \delta\mu/\delta_1$ in the expression

$$\int \frac{E(N^*|\mu)}{N_f^*}\,\delta_1{}^{-1}\,\varphi(\mu/\delta_1)\,d\mu$$

we obtain

$$\int \frac{E(N^*|\mu'\delta_1/\delta)}{N_f^*}\,\delta^{-1}\,\varphi(\mu'/\delta)\,d\mu'.$$

From Lemma 5.3, it follows that,

$$\frac{\int E(N^*|\mu)\,\delta_1{}^{-1}\,\varphi(\mu/\delta_1)\,d\mu}{N_f^*} = \frac{\int E(N|\mu')\,\delta^{-1}\,\varphi(\mu'/\delta)\,d\mu'}{N_f}.$$

6. Bayesian Sequential Methods.

6.1 Introduction.

So far we have only considered frequentist sequential designs. Many non-frequentists have criticised sequential analysis because it contravenes the likelihood principle and relies on a strict adherence to a rather inflexible stopping rule. A detailed outline of these criticisms is given in §6.2. In §6.3 we describe some sequential procedures proposed by Bayesians and Bayesian decision theorists and consider the frequentist properties of one of these designs. We also outline the main objections to non-frequentist approaches. In §6.4 we give a frequentist defence of sequential analysis.

6.2 The Case Against Sequential Analysis.

All the tests considered so far in this thesis have conformed to the frequentist or classical school of statistical inference. In line with the Neyman-Pearson approach to hypothesis testing these tests have been designed to satisfy the Type I and Type II error constraints. Although a Bayesian decision theory approach was employed in §§ 3.6, 4.6 and 5.6 this was not for its own sake, but rather to facilitate an efficient method for the computation of optimal frequentist tests.

The debate between different schools of statistical thought has been particularly vociferous in the area of sequential analysis. The main debate is between **conditionalists** and **unconditionalists**. Conditionalists conduct their analyses conditional on the data observed but unconditionally over the parameter space. Bayesian statisticians are conditionalists. Conversely, unconditionalists conduct their analyses conditional on parameter values but unconditionally over the sample space. Frequentist statisticians are unconditionalists.

Cornfield (1966) defined a **sequential trial** to be '... any form of data collection in which the decision to continue or discontinue further collection depends in some sense on the information previously obtained' and **sequential analysis** to be '... any form of analysis in which the conclusion depends not only on the data, but also on the stopping rule.' Conditionalists and unconditionalists are agreed on the economic and ethical advantages of sequential trials (described in §3.2). They disagree, however, on the relative merits of sequential analysis

with some conditionalists going as far as Anscombe (1963) in believing that ' "Sequential analysis" is a hoax'.

At the centre of the argument is the likelihood principle which states that inferences should be based on the data observed through the likelihood function alone. Conditionalists accept the likelihood principle. For Bayesians, for instance, the principle is implicit in Bayes theorem which states that the posterior distribution for the parameter of interest is proportional to the product of the likelihood function and the prior distribution. There also exist many non-Bayesian arguments in support of the likelihood principle; Cornfield (1966) referenced many of these arguments, while Anscombe (1963) stated that '... the arguments in favour of the principle are indeed weighty'.

Berry (1987) gave two consequences of the likelihood principle:
(i) Data which was not observed should not affect inferences;
(ii) Experiments which were not conducted are irrelevant to inferences.
Much of frequentist inference runs contrary to these consequences and sequential analysis is no exception. To demonstrate this point we consider an example:
Suppose we conduct a two-sided hypothesis test on the mean of a normal distribution (c.f. §4.2). Let $X_1, X_2, \ldots \sim N(\mu,1)$ and suppose we wish test

$$H_1^-: \mu<0 \quad vs \quad H_0: \mu=0 \quad vs \quad H_1^+: \mu>0$$

with Type I error rate $\alpha = 0.05$ and Type II error rate $\beta = 0.1$ at $\mu = \pm 0.25$. Given that S_N $(= X_1 + \ldots + X_N)$ equals $2.1\sqrt{N}$ is there significant evidence to reject the null hypothesis in favour of the alternative?

As frequentists we need further information concerning the stopping rule employed before being able to answer this question. For instance, was a fixed sample size design or a group sequential one used? If the design was group sequential how many interim analyses were planned and what were the nominal significance levels? As we saw in §4.4 the answers to these questions may lead us to radically different conclusions concerning the parameter μ.

Many conditionalists have strongly attacked the reliance of inferences on the stopping rule in sequential analysis. Cornfield (1966), for example, has said that '... to most scientists without previous exposure to statistics, as well as to most intelligent laymen, any dependence of conclusions on stopping rules ... seems like a violation of common sense.' Discussing a similar example Berry (1987)

remarked that '... it flies in the face of what is generally perceived of as science.'

The emphasis on significance levels in our example demonstrates that sequential analysis is inconsistent with consequence (i) of the likelihood principle. For a significance level is the probability under the null hypothesis of observing something at least as extreme as the observed test statistic (see § 7.4) and hence is dependent on what was not observed in addition to what was. Anscombe (1963) criticized significance levels as being '... strictly irrelevant and possibly misleading' in the context of a clinical trial. He argued further that '... consequent inferences, beliefs and actions will be perhaps much affected by what was *not* observed, by all the rest of the sample space besides the point in it' and concluded that '... absurdity can (and in the present case certainly will sometimes) result.'

Conditionalist criticism of sequential analysis is not confined to its contravention of the likelihood principle. On a less philosophical, more practical, level Berry (1987) and Freedman & Spiegelhalter (1989) have attacked the inflexibility of sequential designs. For example suppose we are conducting a clinical trial and we cross a stopping boundary while there are still some patients entered on to the experiment who have not yet responded. What should be done with the data which will ultimately result from these patients? Frequentists following their strict stopping rules cannot use this data directly. Bayesians, on the other hand, can easily incorporate these additional responses into their posterior distributions and therefore their analyses.

Further the decision to stop a clinical trial is a complex one and, although a difference in treatments is obviously important, other factors have to be taken into account. Berry (1985), Berger & Berry (1988) and Freedman & Spiegelhalter (1989) have wondered what a frequentist would do in the event of an experiment being stopped because, for example, one of the treatments is causing undesirable side-effects or because a new treatment with a seemingly better success rate has been developed. The Bayesian can take the data from such an experiment at face value, the frequentist, however, is faced with a problem because his stopping rule has been infringed.

Berry (1985), (1987) has pointed to the difficulty many non-statisticians experience in understanding significance levels. He suggested that many incorrectly regard a P-value as the probability that H_0 is true. This is worrying, for as Berger & Delampady (1987) have shown for a fixed sample size test, even

with a Bayesian analysis based on a noninformative prior distribution the probability of H_0 being true may exceed 0.5 when the corresponding P-value is smaller than 0.05. Berry (1987) also pointed out that many researchers have trouble in understanding why it is necessary to adjust critical values in order to compensate for multiple looks at the data. Accordingly they often do **not** adjust critical values. He argued that such an informal approach to interim analyses is more in line with the Bayesian approach where the data may be analysed as often as one likes without affecting any inferences made. Berry (1987) called for the '... eradication of P-values and significance tests from statistics' and the adoption of more flexible and easily understood Bayesian methods.

Geller & Pocock (1987), Hughes & Pocock (1988) and Freedman & Spiegelhalter (1989) have pointed to the difficulty the frequentist experiences in calculating an unbiased point estimate for the parameter of interest following a sequential experiment. Geller & Pocock also pointed to the difficulty in constructing confidence intervals. A review of the literature on estimation following a sequential experiment is given in § 7.4. We note here, however, that a Bayesian can easily obtain point estimates and confidence intervals from his or her posterior distribution.

Geller & Pocock (1987) listed a number of unresolved problems in sequential analysis. These include how to proceed when more than two treatments are to be compared and how to specify a single clinical endpoint of interest. They also discussed other problems relating to the inflexibility of sequential stopping rules such as how to conduct an unplanned or a retrospective interim analysis.

We also note two often heard criticisms of sequential analysis and frequentist inference in general: the arbitrariness involved in choosing the error rates, α and β, and the failure to incorporate prior information into any analysis even when a large body of data may be available from Phase I and Phase II trials and previous studies.

Many frequentists, notably Armitage (1963), have hit back at the main criticisms expressed here. We shall give a defence of sequential analysis in §6.4.

6.3 Bayesian and Bayesian Decision Theoretic Approaches.

A number of Bayesian and Bayesian decision theoretic approaches have been proposed in the literature. Here we describe 3 of these approaches and consider the frequentist properties of one of them.

Mehta & Cain (1984), Berry (1985) and Freedman & Spiegelhalter (1989) have suggested similar procedures where the decision to stop or to continue at each analysis is made on the basis of the current posterior distribution for the parameter of interest. For example, consider a clinical trial to compare the relative efficacies of two treatments, A and B. A maximum of K groups of n pairs of patients are entered on to the trial with the treatments being randomly assigned within each pair. Suppose X_j represents the difference in response between treatment A and treatment B for the jth pair and that $X_j \sim N(\mu,\sigma^2)$ with σ^2 known. Further let the *a priori* distribution of μ be normal with mean μ_0 and variance ${\sigma_0}^2$. Some simple algebra gives the current distribution of μ at stage i which is also normal with mean

$$\mu_i = \frac{\sigma^2 \mu_0 + {\sigma_0}^2 ni \sum_{j=1}^{ni} X_j}{\sigma^2 + {\sigma_0}^2 ni}$$

and variance

$${\sigma_i}^2 = \frac{\sigma^2 {\sigma_0}^2}{\sigma^2 + {\sigma_0}^2 ni} .$$

The stopping rule of Freedman & Spiegelhalter (1989) is of the general form: At analysis i ($i = 1, 2, \ldots, K$), we stop sampling and conclude that treatment A is superior if

$$\Pr(\mu > \Delta_A | X_1, \ldots, X_{in}) > 1 - \varepsilon. \qquad (6.3.1)$$

Similarly we stop sampling and conclude that treatment B is superior if

$$\Pr(\mu < \Delta_B | X_1, \ldots, X_{in}) > 1 - \varepsilon. \qquad (6.3.2)$$

Here ε is suitably small (for example, $\varepsilon = 0.05$ or 0.025) and (Δ_B, Δ_A) is the range of treatment differences for which the experimenter considers A and B to be equivalent. (Freedman & Spiegelhalter (1983) described a method for eliciting Δ_A and Δ_B from clinicians).

If, for $i < K$, neither inequality (6.3.1) or (6.3.2) is satisfied, the next group of n pairs of patients is entered on to the experiment. The trial is terminated at the Kth analysis with the posterior distribution for μ summing up the experimenter's current beliefs about the actual treatment difference.

It is easily shown that the stopping rule of Freedman & Spiegelhalter is equivalent to stopping to conclude that treatment A is superior when

$$S_{in} = \sum_{j=1}^{ni} X_j > \Delta_A \left[1+\frac{\sigma^2}{ni{\sigma_0}^2}\right] - \frac{\sigma^2\mu_0}{ni{\sigma_0}^2} + \frac{\Phi^{-1}(1-\varepsilon)}{ni{\sigma_0}^2}\sqrt{\sigma^2{\sigma_0}^2(\sigma^2+ni{\sigma_0}^2)} \qquad (6.3.3)$$

and stopping to conclude that treatment B is superior when

$$S_{in} = \sum_{j=1}^{ni} X_j < \Delta_B \left[1+\frac{\sigma^2}{ni{\sigma_0}^2}\right] - \frac{\sigma^2\mu_0}{ni{\sigma_0}^2} - \frac{\Phi^{-1}(1-\varepsilon)}{ni{\sigma_0}^2}\sqrt{\sigma^2{\sigma_0}^2(\sigma^2+ni{\sigma_0}^2)}\,. \qquad (6.3.4)$$

As can be seen, these boundaries are symmetric about zero when $\Delta_B = -\Delta_A$ and $\mu_0 = 0$.

The test proposed by Berry (1985) is a special case of the above method with $\Delta_A = \Delta_B = 0$ and prior mean $\mu_0 = 0$. From (6.3.3) and (6.3.4), the Berry test stops to conclude that treatment A is superior when

$$S_{in} > \frac{\Phi^{-1}(1-\varepsilon)}{ni{\sigma_0}^2}\sqrt{\sigma^2{\sigma_0}^2(\sigma^2+ni{\sigma_0}^2)}$$

and stops to conclude that treatment B is superior when

$$S_{in} < -\frac{\Phi^{-1}(1-\varepsilon)}{ni{\sigma_0}^2}\sqrt{\sigma^2{\sigma_0}^2(\sigma^2+ni{\sigma_0}^2)}.$$

Following McPherson (1982), Berry considered ${\sigma_0}^2$ equal to 0.01 and 0.04.

It is interesting to consider the frequentist properties of the above Bayesian procedure. We concentrate on 4 designs considered by Freedman & Spiegelhalter (1989), with $K = 5$, $n = 20$, $\sigma^2 = 0.5$ $\Delta_A = \Delta_B = 0$, $\varepsilon = 0.025$ and ${\sigma_0}^2 = \sigma^2/n_0$, where $n_0 = 8$, 22, 89 and 800. The different values of n_0 in each design correspond to differing prior beliefs about μ. For example with $n_0 = 8$ these beliefs are diffuse while with $n_0 = 800$ the beliefs are quite concentrated around $\mu = 0$. Table 6.1 gives the Type I error, α, the Type II error at $\mu = \pm 0.25$, β, and the expected sample sizes under $\mu = 0$ and $\mu = \pm 0.25$, with their corresponding minima given underneath in parentheses, for the 4 designs. Also displayed is the working value of ε, $\varepsilon_{0.05}$, which would be necessary to give a test of size 0.05 for each example.

Table 6.1. The Type I error, $\tilde{\alpha}$, Type II error at $\mu = \pm 0.25$, $\tilde{\beta}$, expected sample sizes under $\mu = 0$ and $\mu = \pm 0.25$, together with their corresponding minima in parentheses, and the value of ε, $\varepsilon_{0.05}$, required to give a test with Type I error 0.05 for the designs of Freedman & Spiegelhalter (1989) with $K = 5$, $n = 20$, $\sigma^2 = 0.5$, $\Delta_A = \Delta_B = 0$, $\mu_0 = 0$, $\varepsilon = 0.025$ and $\sigma_0^2 = \sigma^2/n_0$, where $n_0 = 8$, 22, 89 and 800.

	n_0			
	8	22	89	800
$\tilde{\alpha}$	0.098	0.059	0.010	0.0
$\tilde{\beta}$	0.052	0.071	0.188	0.990
$E(N\|\mu=0)$	95.72	97.99	99.84	100.0
	(95.31)	(97.72)	(99.81)	_
$E(N\|\mu=\pm 0.25)$	51.03	58.51	77.94	99.99
	(50.75)	(57.89)	(77.12)	_
$\varepsilon_{0.05}$	0.013	0.022	0.061	0.247

For $n_0 = 800$ the minimum expected sample sizes under $\mu = 0$ and $\mu = \pm 0.25$ were not calculated.

From a frequentist perspective Table 6.1 shows some worrying features. Designs with n_0 "small" (σ_0^2 "large") give rise to tests with narrow boundaries and, consequently, high Type I errors. For example when $n_0 = 8$ ($\sigma_0^2 = 0.0625$) the Type I error is $\tilde{\alpha} = 0.098$. Obviously for smaller values of n_0, $\tilde{\alpha}$ will be even larger.

Conversely designs with n_0 "large" (σ_0^2 "small") give rise to tests with wide boundaries, low Type I errors and high Type II errors. For example, when $n_0 = 800$ ($\sigma_0^2 = 6.25 \times 10^{-4}$) it is almost certain that we will accept the null

hypothesis regardless of the actual treatment difference. Of course with such a precise prior distribution it is unlikely that an experimenter would ever conduct a clinical trial.

For certain choices of n_0, however, it is possible to obtain a test with acceptable size and power under $\mu = \pm 0.25$.

The tests of Freedman & Spiegelhalter (1989) are all very close to optimal in terms of minimizing the expected sample size under $\mu = 0$ and $\mu = \pm 0.25$. Indeed they are never more than one percent of the fixed sample size from the overall minima. Of course, if we compared the expected sample sizes under $\mu = 0$ with the minima for inner wedge tests the results would be less impressive. A Bayesian inner wedge test could be constructed by stating a range of equivalence (Δ_B, Δ_A) and stopping early to accept H_0 if

$$\Pr(\Delta_B < \mu_i < \Delta_A \mid X_1, \ldots, X_{in}) > 1-\varepsilon$$

for suitably small ε. Not only would such an approach improve the efficiency of the Bayes procedure under $\mu = 0$, but it may lead to designs like the one with $n_0 = 800$ being abandoned before any patients are admitted on to the trial.

The final row of Table 6.1 shows the value of ε, $\varepsilon_{0.05}$, required to give a test of size $\alpha = 0.05$. For instance, when $n_0 = 22$, $\varepsilon_{0.05} = 0.022$. However when $n_0 = 800$ it is necessary to consider $\varepsilon_{0.05}$ as large as 0.247 in order to compensate for the concentrated prior distribution. It is very interesting to note that to obtain a test with Type I error 0.05 and power 0.9 at $\mu = \pm 0.25$ it would be necessary to choose $n_0 = 1.786$ (or ${\sigma_0}^2 = 0.28$) and $\varepsilon_{0.05} = 0.009$.

Freedman & Spiegelhalter only considered designs with $K = 5$. A number of frequentists, most notably Armitage (1963), have questioned the effect on error rates of increasing K without adjusting critical values. Table 6.2 gives $\tilde{\alpha}$ and $\tilde{\beta}$ for the 4 designs of Freedman & Spiegelhalter with $K = 5$, 10, 50 and 100.

Table 6.2. The Type I error, $\tilde{\alpha}$, and the Type II error at $\mu = \pm 0.25$, $\tilde{\beta}$, for the designs of Freedman & Spiegelhalter (1989) with $nK = 100$, $\Delta_A = \Delta_B = 0$, $\mu_0 = 0$, $\sigma^2 = 0.5$, $\varepsilon = 0.025$, $n_0 = 8$, 22, 89 and 800 and $K = 5$, 10, 50 and 100.

		n_0			
	K	8	22	89	800
$\tilde{\alpha}$	5	0.098	0.059	0.010	0.0
	10	0.122	0.071	0.017	0.0
	50	0.166	0.095	0.015	0.0
	100	0.179	0.101	0.016	0.0
$\tilde{\beta}$	5	0.052	0.071	0.188	0.990
	10	0.046	0.065	0.179	0.990
	50	0.038	0.055	0.162	0.989
	100	0.036	0.052	0.157	0.989

The pattern that emerges from Table 6.2 is clear; increasing K without adjusting the critical values of the stopping rule increases both the size and power of any given test. Frequentist concerns about sampling to a forgone conclusion (see Armitage et al. (1969), Armitage & McPherson (1971)) are supported by Table 6.2. For example, with $n_0 = 8$, the probability of making an error of the first kind has increased to 0.179 with 100 analyses. Cornfield (1966) suggested that one way of overcoming these problems is to assign a mass of prior probability at $\mu = 0$ and $\mu = \pm 0.25$. Freedman & Spiegelhalter (1989) do not discuss this option.

In addition to the possibility of sampling to a forgone conclusion certain other criticisms have been voiced over these Bayesian designs. Jennison & Turnbull (1989), in response to the discussion of their read paper on repeated confidence intervals, have expressed dissatisfaction at the fact that the designs require no explicit specification of the ethical and economic costs of admitting patients on to a clinical trial. Jennison & Turnbull were also unconvinced about

the wiseness of specifying a prior distribution for the parameter of interest. They pointed to a series of studies conducted by Gilbert, McPeek & Mosteller (1977) where the prior feelings of the clinicians were too optimistic about the efficacy of the experimental treatment. Even with the example discussed above where a normal prior with zero mean was used, a degree of subjectivity is still necessary in specifying the prior variance.

The second approach we consider was first proposed by Anscombe (1963) as an alternative to the frequentist designs of Armitage (1960). Anscombe considered the usual model with n pairs of patients available for comparing the relative efficacies of two treatments, A and B. The difference in response between the ith patient on treatment A and the ith patient on treatment B is denoted by X_i $(i = 1, \ldots, n)$. The X_i's are assumed to be independent and normally distributed with unknown mean μ and unit variance. Following a Bayesian approach, μ is assigned an improper uniform prior distribution.

The Anscombe approach differs from those considered so far in that it assumes that a total of N patients will have their course of treatment affected by the trial - either by being entered on to the experiment or by having their treatment regimen determined by the results of the trial. After the n pairs of patients on the trial have responded we can identify two sources of loss:
i) that arising from n patients receiving the inferior treatment in the clinical trial;
ii) that arising from a wrong decision being made at the conclusion of the trial and $N-2n$ patients receiving the inferior treatment.
Anscombe argued that both types of loss should be proportional to the actual treatment difference μ. Letting $x = X_1 + X_2 + \ldots + X_n$ he suggested that loss (i) should equal $n\,E[|\mu|]$ and loss (ii) should equal $(N-2n)\,E[\max(0,-sgn(x)\mu)]$. Then, as a function of n and x, the total expected loss, $R(n,x)$, equals

$$|x| + \frac{N}{\sqrt{n}}\left\{\varphi\left[\frac{x}{\sqrt{n}}\right] - \frac{|x|}{\sqrt{n}}\Phi\left[-\frac{|x|}{\sqrt{n}}\right]\right\}$$

and the aim is to minimize this loss. Anscombe claimed that to obtain an optimal Bayesian solution to this problem would be '... possible in principle, if difficult in practice, requiring a formidable "backwards induction"' and he suggested an easy to compute near optimal procedure as an alternative. Day (1969) and Chernoff & Petkau (1981) have both used dynamic programming to compute the Bayes test for this problem and for a group sequential analogy respectively.

Frequentists have attacked the Anscombe model because of the difficulty involved in specifying N and because of the use of a prior probability distribution.

A third approach is due to Berry & Ho (1988). They considered using Bayesian decision theory to obtain a procedure for deciding between a new treatment and a control. Their set up is rather different from those we have considered so far in that acceptance of the new treatment is only allowed at the final stage. The motivation for such a design is that when early results point to the superiority of the new treatment testing should continue in order to monitor its safety. In §7.3 we will outline the frequentist version of this problem and extend our method for the computation of optimal tests to cover it. We now give a brief summary of Berry & Ho's method:

As with earlier problems suppose we have a maximum of K groups of n pairs of patients and that X_i $(i = 1, 2, \ldots, nK)$ denotes the difference in response between the patient on the new treatment and the patient on the standard in pair i. It is assumed that the X_i's are independent and normally distributed with unknown mean μ and known variance σ^2 and we wish to decide between

$$D_0: \mu \leq 0 \quad \text{and} \quad D_1: \mu > 0.$$

Berry & Ho assumed that *a priori* μ is normally distributed with mean ∂_0 and variance ${\tau_0}^2$. They further assumed that a unit sampling cost is incurred for each patient entered on to the trial. They suggested a suitable loss function was given by: $L(D_0, \mu) = 0$ with $L(D_1, \mu) = -M\mu$ for $\mu \geq 0$ and $L(D_1; \mu) = L$ for $\mu < 0$. Here M and L are positive constants under the control of the experimenter. The Bayes test for this Bayesian decision problem is computed by dynamic programming (c.f. §3.8). Further details of the dynamic programming algorithm are given in Berry & Ho (1988).

Frequentist criticisms of this procedure concern the difficulties involved in eliciting prior distributions and loss functions from clinicians. To define a loss function over the parameter space would seem to be a very complicated task. It is difficult to see how such a loss function could be arrived at without the introduction of some arbitrariness.

6.4 A Defence of Sequential Analysis.

In §6.3 we gave a number of frequentist criticisms of other inferential approaches to sequential experimentation. We now proceed to defend frequentist sequential analysis and, in particular, our optimal group sequential tests. During the course of this defence we will address many of the Bayesian criticisms outlined in §6.2.

One of the main conditionalist criticisms of sequential analysis in §6.2 concerned its contravention of the likelihood principle. It is worth underlining the point made in §§ 3, 4 and 5 that our optimal group sequential tests are not only feasible and efficient from a frequentist perspective, but also Bayes tests for a suitable choice of prior distribution, cost function and loss function. Moreover, the decision to stop or continue sampling at each stage for the Bayes tests is made on the basis of the current distribution for the parameter of interest. It follows that the Bayes tests are in accordance with the likelihood principle.

Many of the other frequentist group sequential tests suggested in the literature are close to Bayes tests for a suitable decision theory problem. Freedman & Spiegelhalter (1989) have shown that, for a suitable choice of prior, the stopping rules for their Bayesian procedures are similar to the two-sided tests of Pocock (1977) and O'Brien & Fleming (1979). Jennison (1990) pointed to the work of Brown, Cohen & Strawdermann (1980) who proved that "admissible" frequentist tests are Bayes tests for a suitable choice of decision problem, and suggested that '... good frequentist and Bayes stopping rules should not be too dissimilar.'

Much was made in §6.2 concerning the inflexibility of frequentist stopping rules. A good deal of recent research has been devoted to producing more flexible procedures. In §3.12 we saw that Lan & DeMets (1983) and Jennison (1987) have proposed methods for dealing with unpredictable group sizes and numbers of analyses. Our own approach, again outlined in §3.12, attempts to preserve optimality while introducing flexibility and is a useful addition to this area.

As was seen in §4.12 Jennison & Turnbull (1984), (1989) have proposed a highly flexible procedure called the Repeated Confidence Interval (RCI) method. The RCI method enables frequentist inferences to be made even when a trial is stopped for reasons other than the crossing of a stopping boundary. Inferences, then, are still valid if the trial is stopped because, for example, one of the treatments shows unforseen side-effects. Further it is possible to alter our

hypotheses during the trial if desired without invalidating inferences. The method is easily generalized to the one-sided testing problem. Jennison & Turnbull based the calculation of their RCIs on the tests of Pocock (1977) and O'Brien & Fleming (1979). A highly efficient as well as flexible procedure could be obtained by basing the calculation of RCIs on an optimal test.

All the tests we have considered so far have been for normal responses with known variance. However simulation studies by both Pocock (1977) and Gould & Pecore (1982) have shown that group sequential tests are quite robust to departures from both the normality and known variance assumptions. In §7.2 we shall consider the generalization of our efficient method for the computation of optimal group sequential tests to other distributions.

Criticisms relating to estimation following a sequential test are genuine. Research continues in this important area. A review of the literature on estimation is given in §7.4.

Conditionalist criticisms concerning the arbitrariness involved in choosing the error rates of a test are, to some extent, valid. However frequentist methods are not alone in containing a level of arbitrariness and we would advocate the use of power calculations at the design stage of any trial in order to decide on an appropriate stopping rule and maximum sample size.

Finally we feel it is unfair to argue that frequentists never make use of strong prior information when it is available. For example Hughes & Pocock (1988) have suggested employing a Bayesian approach to calculate point estimates for the parameter of interest following a sequential trial. Also Jennison (1986) advocated the use of prior distributions in the design of optimal frequentist tests for the bioequivalence problem (see §5.11).

It is perhaps appropriate to end this section with a quote from Anscombe (1963) who is very critical of the tests proposed in the book by Armitage (1960), but states that the book's '... net effect on medical research will almost certainly be good, ... partly because any consideration of sequential designs encourages flexibility.'

7. Further Work and Other Topics.

7.1 Introduction.

In §7 we consider some suggestions for further work as well as briefly mentioning some topics in sequential analysis which have not been looked at so far.

In §7.2 we consider the extension of our efficient method for the computation of optimal group sequential tests to non-normal responses. In Phase II clinical trials it is usually appropriate to consider Bernoulli responses to treatments and we propose an approach for computing optimal tests for this problem.

In §7.3 we look at the problem considered by Berry & Ho (1988) where a new treatment is tested against a standard with early stopping only being permitted for the rejection of the new treatment. Berry & Ho considered this problem from a Bayesian perspective. We propose extending our method of §§ 3.6-3.8 to obtain an efficient method for computing optimal frequentist tests for this problem.

The calculation of P-values, confidence intervals and point estimates following a sequential trial is considered in §7.4. Finally, in §7.5, we review the literature on stochastic curtailment and its Bayesian counterpart based on predictive power.

7.2 Optimal Group Sequential Tests for Non-Normal Data.

So far we have only considered group sequential tests for normal data. With group sequential procedures we deal with sums of random variables and so we might expect our tests to be robust to departures from normality (by the Central Limit Theorem) and the known variance assumption. Simulation studies conducted by Pocock (1977) and Gould & Pecore (1982) have demonstrated this robustness. Another possibility, suggested by Armitage (1975), is to transform non-normal data by taking logs, for example, to facilitate the use of tests designed for normal responses.

We have not carried out any studies into the robustness of our own optimal group sequential tests. However, on the basis of the work of both Pocock and Gould & Pecore, it would seem reasonable to assume that they are fairly robust.

Our method for computing optimal group sequential tests is easily generalized to cases where the data is non-normal. Of particular interest are optimal group sequential tests for the binomial distribution. Fleming (1982) and Chang, Therneau, Wishart & Cha (1987) considered Phase II trials for anticancer drugs. The emphasis in Phase II trials is on screening; we wish to decide whether a drug is ineffective and should be discarded or whether it merits further investigation in the form of a Phase III trial. In the case of anticancer drugs interest typically centres on the parameter p termed the "regression probability" which, according to Chang *et al.*, is '... the probability an eligible patient receiving the treatment will experience a tumour regression as precisely defined in the protocol'.

Following the notation of Chang *et al.* (1987) we suppose that a maximum of N patients are available for entry on to our Phase II trial. Further let X_i ($i = 1,\ldots,N$) be the random variable representing the response of the ith patient. Then the X_i's are independent Bernoulli random variables with success probability p ($0 \leq p \leq 1$), i.e.

$$\Pr(X_i = 1) = p, \quad \Pr(X_i = 0) = 1-p.$$

We consider testing

$$H_0 : p = p_0 \quad vs \quad H_1 : p = p_1$$

with error probabilities

$$\Pr(\mathcal{A}_1 \mid p = p_0) = \alpha \tag{7.2.1}$$

$$\Pr(\mathcal{A}_0 \mid p = p_1) = \beta \tag{7.2.2}$$

Here $p_0 < p_1$, and $\mathcal{A}_0$ and $\mathcal{A}_1$ denote the acceptance of H_0 and H_1 respectively.

Acceptance of H_0 would lead us to abandon our investigations into the drug, while rejection of H_0 in favour of H_1 would lead to further studies into the drug's efficacy. Because of the discrete nature of the binomial distribution it will not, in general, be possible to obtain stopping rules satisfying (7.2.1) and (7.2.2) exactly. Chang *et al.* suggested that, instead, we should consider tests with Type I error at most α and Type II error at most β. We shall term such tests "feasible".

The economic and ethical arguments put forward in § 3.2 for introducing group sequential designs into Phase III clinical trials are equally valid for Phase II trials. So we shall consider a maximum of K groups of sizes $n_1, n_2 - n_1, \ldots, n_K - n_{K-1}$ for the above testing problem and stopping rules of the form:

At analysis i $(1 \le i \le K-1)$,

$$\text{if } \sum_{j=1}^{n_i} X_j \ge b_i \text{ stop sampling and accept } H_1;$$

$$\text{if } \sum_{j=1}^{n_i} X_j \le a_i \text{ stop sampling and accept } H_0;$$

$$\text{if } a_i < \sum_{j=1}^{n_i} X_j < b_i \text{ continue sampling.}$$

At analysis K,

$$\text{if } \sum_{j=1}^{n_K} X_j \ge a_K + 1 \text{ stop sampling and accept } H_1;$$

$$\text{if } \sum_{j=1}^{n_K} X_j \le a_K \text{ stop sampling and accept } H_0.$$

Here $-1 \le a_i \le b_i \le n_i + 1$ $(i = 1, \ldots, K)$.

The frequentist properties of our stopping rule may be calculated numerically. Equations (7.2.3), (7.2.4) and (7.2.5) give the Type I error, $\tilde{\alpha}$, Type II error, $\tilde{\beta}$, and expected sample size under p, $E(N|p)$, respectively. To ease notation we let m_j denote the number of observations in the jth group, i.e. $m_j = n_j - n_{j-1}$ $(j = 2, \ldots, K)$ and $m_1 = n_1$.

$$\tilde{\alpha} = \sum_{j=1}^{K} \{ \sum_{i_j=b_j}^{n_j} \sum_{i_{j-1}=a_{j-1}+1}^{b_{j-1}-1} \cdots \sum_{i_1=a_1+1}^{b_1-1} \Pr_{p_0}(S_{n_j} - S_{n_{j-1}} = i_j - i_{j-1}) \ldots \Pr_{p_0}(S_{n_1} = i_1) \} \quad (7.2.3)$$

$$\tilde{\beta} = \sum_{j=1}^{K} \{ \sum_{i_j=a_{j-1}+1}^{a_j} \sum_{i_{j-1}=a_{j-1}+1}^{b_{j-1}-1} \cdots \sum_{i_1=a_1+1}^{b_1-1} \Pr_{p_1}(S_{n_j} - S_{n_{j-1}} = i_j - i_{j-1}) \ldots \Pr_{p_1}(S_{n_1} = i_1) \} \quad (7.2.4)$$

$$E(N|p) = \sum_{j=1}^{K} \{ \sum_{i_j \in r_j} \sum_{i_{j-1}=a_{j-1}+1}^{b_{j-1}-1} \cdots \sum_{i_1=a_1+1}^{b_1-1} \Pr_{p}(S_{n_j} - S_{n_{j-1}} = i_j - i_{j-1}) \ldots \Pr_{p}(S_{n_1} = i_1) \} \quad (7.2.5)$$

Here $r_j = \{(0, a_j) \cup (b_j, n_j)\}$, for $j = 1, 2, \ldots, K$, and the distribution of $S_{n_j} - S_{n_{j-1}}$ $(j = 2, \ldots, K)$ is binomial with parameters m_j and p.

Chang *et al.* (1987) considered the computation of optimal group sequential binomial designs. For instance they considered computing the feasible test which minimizes $E(N|p_0) + E(N|p_1)$. They noted that the so called exhaustive

approach to this problem (which considers all possible designs) is both computationally inefficient and expensive. For example (see Chang *et al.* for further details) with $p_0 = 0.3$, $p_1 = 0.5$, $K = 3$, $n_1 = 20$, $n_2 = 15$, $n_3 = 15$, $\alpha = 0.05$ and $\beta = 0.2$, over 300 000 designs have to be considered. Chang *et al.* laid down a number of constraints which must be satisfied by the optimal test, and by so doing reduced the number of tests that need be considered to just 15 000.

Clearly it is possible to extend our efficient method for the computation of optimal group sequential tests for normally distributed data to binomially distributed data. This is achieved by considering the following family of problems in Bayesian decision theory:
Suppose there are a maximum of K groups of sizes $m_1, m_2, \ldots, m_K$ of independent Bernoulli random variables with success parameter p available for making a decision between:

$$D_0 : p = p_0 \quad \text{and} \quad D_1 : p = p_1 .$$

The common prior distribution for p is given by $\pi(p_0) = \pi(p_1) = 1/2$ with $\pi(p) = 0$ otherwise, while the common cost of sampling function is given by $c(p_0) = c(p_1) = 1$ with $c(p) = 0$ otherwise. Individual problems within the family differ in their loss functions, $L(D,p)$, which are indexed by a pair of parameters, d_0 (>0) and $d_1(>0)$. The general form of $L(D,p)$ is given by : $L(D_1,p_0) = d_0$ and $L(D_0,p_1) = d_1$, with $L(D,p) = 0$ otherwise.

Suppose for the moment that d_0 and d_1 are fixed and consider a general decision rule for the above problem, $\mathcal{B}$. The general form of $\mathcal{B}$ is given by:

At analysis i $(1 \le i \le K-1)$,

$$\text{if } \sum_{j=1}^{n_i} X_j \ge b_i \text{ stop sampling and make decision } D_1$$

$$\text{if } \sum_{j=1}^{n_i} X_j \le a_i \text{ stop sampling and make decision } D_0$$

$$\text{if } a_i < \sum_{j=1}^{n_i} X_j < b_i \text{ continue sampling.}$$

At analysis K,

$$\text{if } \sum_{j=1}^{n_K} X_j \ge a_K + 1 \text{ stop sampling and make decision } D_1$$

$$\text{if } \sum_{j=1}^{n_K} X_j \le a_K \text{ stop sampling and make decision } D_0.$$

The **risk** of $\mathcal{B}$, $r(\mathcal{B},d_0,d_1)$, is defined as the sum of the total expected sampling cost **plus** the total expected loss through making a wrong decision. That is,

$$r(\mathcal{B},d_0,d_1) = c(p_0)\,E(N|p=p_0)\,\pi(p_0) + c(p_1)\,E(N|p=p_1)\,\pi(p_1)$$
$$+ d_0\,\Pr(\mathcal{A}_0|p=p_1)\,\pi(p_1) + d_1\,\Pr(\mathcal{A}_1|p=p_0)\,\pi(p_0)$$

The Bayes decision rule for our problem, $\mathcal{B}^*(d_0,d_1)$, minimizes this risk over the set of all decision rules. We can compute $\mathcal{B}^*(d_0,d_1)$ by dynamic programming. (We omit details of the dynamic programming algorithm, which is a discrete analogy of that for normally distributed data.) Using numerical integration we can calculate the errors of $\mathcal{B}^*(d_0,d_1)$. Suppose these errors are given by

$$\Pr(D_1|p=p_0) = \tilde{\alpha}$$

and

$$\Pr(D_0|p=p_1) = \tilde{\beta},$$

then clearly

$$r(\mathcal{B}^*(d_0,d_1),d_0,d_1) = \frac{1}{2}\left\{ E(N|p=p_0) + E(N|p=p_1) + d_0\,\tilde{\alpha} + d_1\,\tilde{\beta} \right\}.$$

It follows from the definition of the Bayes rule that there can be no other decision rule with errors $\tilde{\alpha}$ and $\tilde{\beta}$ which attains a lower value of $E(N|p=p_0) + E(N|p=p_1)$. By searching over (d_0,d_1) we obtain the feasible test which minimizes $E(N|p=p_0) + E(N|p=p_1)$.

The main problem with this approach concerns the search over (d_0,d_1) to obtain a loss function leading to a Bayes test with the required errors. The discreteness inherent in our problem means that if we consider a plot of $\tilde{\alpha}$ as a function of d_0,d_1 we observe a 2-dimensional step function. Similarly a plot of $\tilde{\beta}$ as a function of d_0 and d_1 would also give a 2-dimensional step function. However we note that if we fix d_0 and increase d_1 then $\tilde{\alpha}$ decreases monotonically. Also if we fix d_1 and increase d_0 then $\tilde{\beta}$ decreases monotonically. Hence the problem of searching over d_0 and d_1 for the Bayes test with the required error rates can be reduced to a series of one-dimensional bisection searches.

As was mentioned earlier, we do not have any results for this proposal. A study of it would be worthwhile in order to determine whether, and if so by how much, it is more efficient than the approach of Chang *et al.* (1987).

7.3 The Problem of Berry & Ho.

Berry & Ho (1988) introduced a Bayesian group sequential procedure for comparing an experimental treatment with a standard. The problem they considered differed from the one in §3 in that early stopping was only permitted for the acceptance of H_0 (i.e. the standard treatment is no worse than the experimental). Rejection of H_0 was only allowed at the final analysis. Such a design is attractive as it attempts to stop trials involving unpromising experimental treatments as quickly as possible, while it allows trials involving promising experimental treatments to continue so that secondary issues may be analysed.

Clearly our improved method for the computation of optimal one-sided group sequential tests given in §3 can be extended to this problem. We shall now describe this extension: Suppose a maximum of K groups of n pairs of patients are available for entry on to a trial and that X_i, the difference in response between the ith patient on the new treatment and that on the standard, is normally distributed with unknown mean μ and known variance σ^2. We wish to test

$$H_0: \mu \leq 0 \quad vs \quad H_1: \mu > 0$$

with error rates

$$\Pr(\mathcal{A}_1 \,|\, \mu = 0) = \alpha \tag{7.3.1}$$

and

$$\Pr(\mathcal{A}_0 \,|\, \mu = \delta) = \beta. \tag{7.3.2}$$

Our stopping rule is of the form:

At analysis i $(1 \leq i \leq K-1)$,

if $S_{in} \leq c_i$ stop entering patients on to the trial and accept H_0;

if $S_{in} > c_i$ enter the next group of n pairs of patients on to the trial.

At analysis K,

if $S_{Kn} \leq c_K$ stop entering patients on to the trial and accept H_0;

if $S_{Kn} > c_K$ stop entering patients on to the trial and accept H_1.

Clearly, under $\mu \leq 0$, we would like to stop the above experiment and accept the null hypothesis as soon after the start of the trial as possible. Consider, for example, the minimization of objective function $F_1 : E(N|\mu=0)$ over feasible stopping rules of the above form. Below we describe how our improved method for the computation of optimal one-sided group sequential tests can be extended to this problem.

Consider the following family of problems in Bayesian decision theory:

For the clinical trials problem described earlier we are interested in choosing between

$$D_0 : \mu \leq 0 \quad \text{and} \quad D_1 : \mu > 0.$$

Our family of problems has the common prior distribution for μ given by: $\pi(0) = \pi(\delta) = 1/2$ with $\pi(\mu) = 0$ otherwise, and the common cost of sampling function is given by: $c(0) = 1$ with $c(\mu) = 0$ otherwise. Individual problems within the family differ in terms of their loss functions, $L(D;\mu)$, which are indexed by a pair of parameters d_0 (> 0) and d_1 (> 0). The general form of $L(D,\mu)$ is given by $L(D_1,0) = d_0$, $L(D_0,\delta) = d_1$ with $L(D;\mu) = 0$ otherwise.

Suppose for the moment that d_0 and d_1 are fixed and consider a general decision rule for this problem, $\mathcal{B}$. The general form of $\mathcal{B}$ is given by:

At analysis i $(1 \leq i \leq K-1)$,

if $S_{in} \leq c_i$ stop entering patients on to the trial and make decision D_0;

if $S_{in} > c_i$ enter the next group of n pairs of patients on to the trial.

At analysis K,

if $S_{Kn} \leq c_K$ stop entering patients on to the trial and make decision D_0;

if $S_{Kn} > c_K$ stop entering patients on to the trial and make decision D_1.

The **risk** of $\mathcal{B}$, $r(\mathcal{B},d_0,d_1)$, is defined as the sum of the total expected sampling cost **plus** the total expected loss through making a wrong decision. That is

$$r(\mathcal{B},d_0,d_1) = c(0)\, E(N|\mu=0)\, \pi(0) + d_0 \Pr(D_1|\mu=0)\, \pi(0) + d_1 \Pr(D_0|\mu=\delta)\, \pi(\delta).$$

The Bayes decision rule for this problem, $\mathcal{B}^*(d_0,d_1)$, minimizes this risk over the set of all decision rules. We can compute $\mathcal{B}^*(d_0,d_1)$ by dynamic programming (details are omitted although, clearly, the algorithm is similar to those described in earlier sections). Using numerical integration we can calculate the errors of $\mathcal{B}^*(d_0,d_1)$. Suppose these errors are given by

$$\Pr(D_1 \mid \mu = 0) = \tilde{\alpha}$$

and

$$\Pr(D_0 \mid \mu = \delta) = \tilde{\beta}$$

then clearly

$$r(\mathcal{B}^*(d_0,d_1),d_0,d_1) = \frac{1}{2}\{F_1 + d_0\,\tilde{\alpha} + d_1\,\tilde{\beta}\}.$$

It follows from the definition of the Bayes rule that there can be no other decision rule with errors $\tilde{\alpha}$ and $\tilde{\beta}$ which attains a lower value of F_1. By searching over (d_0,d_1) we obtain a pair of loss parameters, $(d_0^{(\alpha)},d_1^{(\beta)})$ giving rise to a Bayes decision theory problem with an associated Bayes rule, $\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)})$, with errors α and β. It follows that

$$r(\mathcal{B}^*(d_0^{(\alpha)},d_1^{(\beta)}),d_0^{(\alpha)},d_1^{(\beta)}) = \frac{1}{2}\{F_1 + d_0^{(\alpha)} + d_1^{(\beta)}\},$$

and from the definition of the Bayes decision rule there can be no other decision rule with errors α and β which attains a lower value of F_1. By equating decisions D_0 and D_1 with stopping to accept the hypotheses H_0 and H_1 we have computed the optimal feasible test for our original frequentist problem.

As was made clear in §6.3 there is an increasing emphasis on designs like the one considered in this section. There is a strong case to be made, therefore, for future research into optimal frequentist tests for this problem. It would also be a useful exercise to compare the attained error rates and expected sample sizes of our optimal tests with those of the tests proposed by Berry & Ho.

7.4 P-Values, Confidence Intervals and Point Estimation Following a Group Sequential Experiment.

This thesis has concentrated on designs for group sequential hypothesis tests. It would be misleading to pretend that at the end of a clinical trial a simple

decision to accept or reject the null hypothesis is all that is required. Point estimates and confidence intervals for the parameter(s) of interest and P-values (also known as observed significance levels) will also be of use in any statistical analysis. In a fixed sample size study inferences are made conditional on the sample size. In a sequential experiment the parameter of interest can substantially influence the final sample size and so it would be inappropriate to make inferences conditional on this sample size. Hence a range of new methods have been developed and proposed in the literature. All of these methods relate to inferences after a one-sided or two-sided hypothesis test.

We begin by considering P-values:

P-values.

A P-value is defined as the probability under the null hypothesis of observing something at least as extreme as the observed test statistic. Hence a P-value is a useful summary of just how likely the observed data is under H_0. In a sequential setting the main problem when calculating P-values comes in defining what is meant by "at least as extreme as the observed test statistic". Suppose our experiment stops at the ith analysis ($1 \le i \le K$) and that $S_{in} = \sum_{j=1}^{in} X_j$ is equal to x.

Then for the two-sided testing problem with no inner wedge Madsen & Fairbanks (1983) defined a P-value to be the total probability under $\mu = 0$ of stopping and rejecting H_0 at analysis j ($j < i$) and of stopping at analysis i with $|S_{in}| \ge x$, i.e.

$$\sum_{j=1}^{i-1} \Pr_0(|S_n| < c_1, \ldots, |S_{(j-1)n}| < c_{j-1}, |S_{jn}| \ge c_j)$$

$$+ \Pr_0(|S_n| < c_1, \ldots, |S_{(i-1)n}| < c_{i-1}, |S_{in}| \ge x).$$

The probabilities in the above expression can be expressed in terms of multiple integrals and calculated numerically (see Appendix 5.1). Of course if $i = K$ and $x = \pm c_K$ the P-value is exactly α.

For one-sided tests we define a P-value to be the probability under $\mu = -\delta$ ($\mu = \delta$) of stopping and rejecting H_1^- (H_1^+) at analysis j ($j < i$) and of

stopping at analysis i with $S_{in} \geq x$, i.e. (under H_1^-) the P-value is given by

$$\sum_{j=1}^{i-1} \Pr_{-\delta}(|S_n| < c_1 , \ldots , |S_{(j-1)n}| < c_{j-1} , S_{jn} \geq c_j)$$

$$+ \Pr_{-\delta}(|S_n| < c_1 , \ldots , |S_{(i-1)n}| < c_{i-1} , S_{in} \geq x).$$

Clearly, if $i = K$ and $x = 0$, the P-value is exactly α.

Confidence Intervals.

A $100(1-\alpha)\%$ confidence interval (CI) for μ is defined to be a set of parameter values within which the true value μ lies with probability $(1-\alpha)$. We note that the calculation of a fixed sample size $100(1-\alpha)\%$ CI following a group sequential experiment is invalid.

Tsiatis, Rosner & Mehta (1984) have suggested the following approach for constructing a $100(1-\alpha)\%$ CI for μ after a sequential test: Suppose our test stops at analysis i and that the corresponding sum of observations is S_{in}. Then we term the pair (i , S_{in}) **the stopping time** of the test. There is obviously a set of possible stopping times $\{(i , S_{in}): 1 \leq i \leq K\}$ and the first stage in constructing a CI is to order these times. Tsiatis *et al.* suggested the following (ascending) order of stopping times:

$$(1,-\infty) , \ldots , (1,-c_1),$$

$$(2,-\infty) , \ldots , (2,-c_2),$$

$$\ldots$$

$$(K,-\infty),\ldots,(K,-c_K) , \ldots , (K,c_K) , \ldots , (K,\infty),$$

$$(K-1,c_{K-1}) , \ldots , (K-1,\infty),$$

$$\ldots$$

$$(1,c_1) , \ldots , (1,\infty).$$

We use the notation $(i,S_{in}) \geq (i',S_{in}')$ to denote the fact that (i,S_{in}) is at least as large as (i',S_{in}'). For an observed stopping time (i',S_{in}') and given values of μ, μ_1 and μ_2, we can calculate

$$\Pr_{\mu_1}((i,S_{in}) > (i',S_{in}'))$$

and

$$\Pr_{\mu_2}((i,S_{in}) < (i',S_{in}')).$$

Using numerical integration and the bisection search method we obtain μ_1 and μ_2 such that the above probabilities both equal $\alpha/2$. Then (μ_1, μ_2) is a $100(1-\alpha)\%$ CI for μ. The above approach may be used for both one-sided and two-sided tests without an inner wedge. The ordering of stopping times when an inner wedge is present has not been discussed in the literature.

We note that Jennison & Turnbull (1983) have suggested a similar approach for constructing confidence intervals for the probability of success parameter of the binomial distribution.

Hughes & Pocock (1988) have suggested a Bayesian approach for obtaining confidence intervals following a frequentist design. They suggested assigning μ a prior distribution at the start of the trial and then calculating its posterior distribution on the basis of the observed data after the experiment has ended. From this posterior distribution a suitable $100(1-\alpha)\%$ CI may be constructed.

Point Estimates.

We saw in §6.2 that an outstanding problem in sequential analysis concerns the calculation of unbiased point estimates for the parameter(s) of interest after an experiment. As an example consider again the two-sided hypothesis test on the mean of the normal distribution, μ. For a fixed sample size experiment the maximum likelihood estimate (m.l.e.) of μ, $\hat{\mu}$, is unbiased. i.e. $E(\hat{\mu}) = \mu$. However the m.l.e. of μ after a sequential experiment may be substantially biased. i.e. $E(\hat{\mu}) = \mu + b(\mu)$, where $b(\mu)$ is the bias function. This is a particularly acute problem when our test stops early to reject H_0.

The reason why m.l.e.'s after a sequential trial are biased concerns the data dependent stopping rule; we stop early because there is strong evidence against H_0.

It is relatively simple to calculate a median unbiased point estimate of μ, $\tilde{\mu}$, after a sequential trial. Median unbiased estimates are such that $med(\tilde{\mu}) = \mu$. Jennison & Turnbull (1989) suggested ordering the stopping times of our test as with the confidence intervals above. A numerical search method is then used to calculate $\tilde{\mu}$ such that

$$\Pr_{\tilde{\mu}}((i, S_{in}) \geq (i', S_{in}')) = 0.5.$$

To compute mean unbiased estimates of μ Whitehead (1986) suggested estimating the bias in $\hat{\mu}$, b($\hat{\mu}$), and then subtracting it from $\hat{\mu}$. He noted that b($\hat{\mu}$) could be obtained by numerical integration after a group sequential trial.

Hughes & Pocock (1988) have suggested a Bayesian method for the calculation of point estimates analogous to their method for CIs. Their claim is that the resulting point estimates, based on a posterior distribution for μ, correct for any bias which may be present.

7.5 Stochastic Curtailment and Predictive Power.

Stochastic curtailment was initially proposed in the literature by Halperin, Lan, Ware, Johnson & DeMets (1982). It is an alternative method to sequential analysis which allows a single sample study to be terminated early if there is a "high" chance that the fixed sample size test is going to accept or reject the null hypothesis.

As an example consider again the fixed sample size one-sided test on the mean of the normal distribution, μ, introduced in §3.2. After N_f' observations we reject H_0: $\mu \leq 0$ if the standardized test statistic, $S_{N_f'}/\sqrt{N_f'\sigma^2}$, is greater than $\Phi^{-1}(1-\alpha)$ and accept H_0 otherwise. A suitable choice of N_f' gives a test with Type II error β at $\mu = \delta$.

Suppose that after n ($<N_f'$) observations the sum of the observed responses equals x. Halperin *et al.* (1982) suggested stopping the trial and rejecting H_0 in favour of H_1 if

$$\Pr_0(S_{N_f'}/\sqrt{N_f'\sigma^2} \geq \Phi^{-1}(1-\alpha) \mid S_n = x) \geq \gamma \qquad (7.5.1)$$

where γ is suitably large. In other words we reject the null hypothesis after n observations if the conditional probability of rejecting it after N observations is "large". The probability on the left hand side of (7.5.1) is easily calculated numerically.

Similarly we accept H_0 if

$$\Pr_\delta(S_{N_f'}/\sqrt{N_f'\sigma^2} < \Phi^{-1}(1-\alpha) \mid S_n = x) \geq \gamma'. \qquad (7.5.2)$$

Halperin *et al.* proved that under the above "stopping rule" the Type I error is bounded above by α/γ and the Type II error is similarly bounded by β/γ'. The

generalization of stochastic curtailment to two-sided tests with or without early acceptance of H_0 is obvious.

The Bayesian alternative to stochastic curtailment is known as **predictive power** and was discussed by Spiegelhalter, Freedman & Blackburn (1986). The decision to stop or continue the trial after n observations is made on the basis of the predicted probability of accepting or rejecting H_0. This probability is **not** conditional on any parameter value(s).

Suppose the prior distribution for μ is denoted by $\pi(\mu)$. Spiegelhalter *et al.* (1986) recommended the use of a noninformative prior - for example a normal prior with zero mean and variance ${\tau_0}^2$. Then the current distribution for μ after n observations is denoted by $\pi^{(n)}(\mu|x)$ and the probability of the test based on N observations rejecting H_0, the predictive power, equals

$$\int_{\mu} \Pr_{\mu}(S_N \geq c_N \mid x)\, \pi^{(n)}(\mu|x)\, d\mu,$$

where c_N is the critical value of the Bayes test. We reject H_0 if the predictive power is sufficiently large and accept H_0 if it is sufficiently small. Again the integral may be calculated numerically.

Both stochastic curtailment and the Bayesian method based on predictive power are important approaches which add a degree of flexibility to experimental designs. An interesting suggestion for further research would be to consider the efficiency of these designs relative to our optimal group sequential tests.

References

Anderson, T.W. (1960). A modification of the sequential probability ratio test to reduce the sample size. *Ann. Math. Statist.* **31**, 165-97.

Anscombe, F.J. (1963). Sequential medical trials. *J. Am. Statist. Assoc.* **58**, 365-83.

Armitage, P. (1957). Restricted sequential procedures. *Biometrika* **44**, 9-26.

Armitage, P. (1960). *Sequential Medical Trials*. Springfield, Illinois: Thomas, (1st. Ed.).

Armitage, P. (1963). Sequential medical trials; some comments on F.J. Anscombe's paper. *J. Am. Statist. Assoc.* **58**, 384-7.

Armitage, P. (1975). *Sequential Medical Trials*. (2nd Edition). Oxford: Blackwell.

Armitage, P., McPherson, K. & Rowe, B.C. (1969). Repeated significance tests on accumulating data. *J. R. Statist. Soc. A* **132**, 235-44.

Berger, J.O. & Berry, D.A. (1988). Statistical analysis and the illusion of objectivity. *The American Scientist* **76**, 159-65.

Berger, J.O. & Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.* **2**, 317-52.

Berry, D.A. (1985). Interim analyses in clinical trials: classical vs. bayesian approaches. *Statistics in Medicine* **4**, 521-6.

Berry, D.A. (1987). Interim analyses in clinical trials: the role of the likelihood principle. *The American Statistician* **41**, 117-22.

Berry, D.A. & Ho, C-H. (1988). One-sided sequential stopping boundaries for clinical trials: a decision-theoretic approach. *Biometrics* **44**, 219-27.

Brown, L.D., Cohen, A. & Strawderman, W.E. (1979). Monotonicity of Bayes sequential tests. *Ann. Statist.* **7**, 1222-30.

Chang, M.N., Therneau, T.M., Wieand, H.S. & Cha, S.S. (1987). Designs for group sequential Phase II clinical trials. *Biometrics* **43**, 865-74.

Chernoff, H. & Petkau, A.J. (1981). Sequential medical trials involving paired data. *Biometrika* **68**, 119-32.

Colton, T. & McPherson, K. (1976). Two-stage plans compared with fixed sample size and Wald SPRT plans. *J. Am. Statist. Assoc.* **71**, 80-86.

Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle. *The American Statistician* **20**, 18-23.

Cuzick, J. (1981). Boundary crossing probabilities for stationary gaussian-processes and Brownian motion. *T. Am. Math. S.* **263**, 469-92.

Day, N.E. (1969). A comparison of some sequential designs. *Biometrika* **56**, 301-11.

DeMets, D.L. (1987). Practical aspects of data monitoring: a brief review. *Statistics in Medicine* **6**, 753-60.

DeMets, D.L. & Ware, J.H. (1980). Group sequential methods in clinical trials with a one-sided hypothesis. *Biometrika* **67**, 651-60.

DeMets, D.L. & Ware, J.H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* **69**, 661-3.

Dodge, H.F. & Romig, H.G. (1929). A method of sampling inspection. *Bell. Syst. Tech. Jrnl.* **8**, 613-31.

Elashoff, J.D. & Reedy, T.J. (1984). Two-stage clinical trial stopping rules. *Biometrics* **40**, 791-5.

Emerson, S.S. & Fleming, T.R. (1989). Symmetric group sequential test designs. *Biometrics* **45**, 905-23.

Fleming, T.R. (1982). One-sample multiple testing procedure for Phase II clinical trials. *Biometrics* **38**, 143-51.

Fleming, T.R., Harrington, D.P. & O'Brien, P.C. (1984). Designs for group sequential clinical tests. *Controlled Clinical Trials* **5**, 348-61.

Freedman, L.S., Lowe, D. & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics* **40**, 575-86.

Freedman, L.S. & Spiegelhalter, D.J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *The Statistician* **33**, 153-60.

Freedman, L.S. & Spiegelhalter, D.J. (1989). Bayesian monitoring of clinical trials. *Controlled Clinical Trials* **10**, 357-67.

Geller, N.L. & Pocock, S.J. (1987). Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. *Biometrics* **43**, 213-23.

Gilbert, J.P., McPeek, B. & Mosteler, F. (1977). Statistics and ethics in surgery and anaesthesia. *Science* **198**, 684-9.

Gore, S.M. (1981). Statistics in question - Assessing clinical trials - Design 1. *British Medical Journal* **282**, 1780-1.

Gould, A.L. (1983). Abandoning lost causes (early termination of unproductive clinical trials). *Proc. Biopharm. Soc. ASA.*

Gould, A.L. & Pecore, V.J. (1982). Group sequential methods for clinical trials allowing early acceptance of H_0 and incorporating costs. *Biometrika* **69**, 75-80.

Halperin, M., Lan, K.K.G., Ware, J.H., Johnson, N.J. & DeMets, D.L. (1982). An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials* **3**, 311-23.

Haybittle, J.L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology* **44**, 793-7.

Hughes, M.D. & Pocock, S.J. (1988). Stopping rules and estimation problems in clinical trials. *Statistics in Medicine* **7**, 1231-42.

Jennison, C. (1986). Discussion on Bayesian methods in practice: experiences in the pharmaceutical industry. (by Racine-Poon *et al.*) *Appl. Statist.* **35**, 131.

Jennison, C. (1987). Efficient group sequential tests with unpredictable group sizes. *Biometrika* **74**, 155-65.

Jennison, C. (1990). Comment on a paper by N. Breslow. *Statist. Sci.* **5**, 288-91.

Jennison, C. & Turnbull, B.W. (1983). Confidence intervals for a binomial parameter following a multistage test with applications to MIL-STD 105D and medical trials. *Technometrics* **25**, 49-58.

Jennison, C. & Turnbull, B.W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials* **5**, 33-45.

Jennison, C. & Turnbull, B.W. (1989). Interim analyses: the repeated confidence interval approach (with discussion). *J. R. Statist. Soc. B* **51**, 305-61.

Johnson, N.L. (1961). Sequential Analysis: a survey. *J. R. Statist. Soc. A* **124**, 372-411.

Lai, T.L. (1973). Optimal stopping and sequential tests which minimize the maximum expected sample size. *Ann. Statist.* **1**, 659-73.

Lan, K.K.G. & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-64.

Lan, K.K.G. & DeMets, D.L. (1989). Changing the frequency of interim analysis in sequential monitoring. *Biometrics* **45,** 1017-20.

Lorden, G. (1976). 2-SPRT's and the modified Keifer-Weiss problem of minimizing an expected sample size. *Ann. Statist.* **4**, 281-91.

Madsen, R.W. & Fairbanks, K.B. (1983). P-values for multistage and sequential tests. *Technometrics* **25**, 285-93.

McPherson, K. (1974). Statistics: the problem of examining accumulating data more than once. *New England Journal of Medicine* **290**, 501-2.

McPherson, K. (1982). On choosing the number of interim analyses in clinical trials. *Statistics in Medicine* **1**, 25-36.

McPherson, K. & Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *J. R. Statist. Soc. A*, **134**, 15-25.

Mehta, C.R. & Cain, K.C. (1984). Charts for early stopping of pilot studies. *Journal of Clinical Oncology* **2**, 676-82.

Metzler, C.M. (1974). Bioavailability - a problem of equivalence. *Biometrics*, **30**, 309-17.

Nelder, J.A. & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal* **7**, 308-13.

O'Brien, P.C. & Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549-56.

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. & Smith, P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observations of each patient. I. Introduction and design. *British Journal of Cancer* **34**, 585-612.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-9.

Pocock, S.J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics* **38**, 153-62.

Pocock, S.J. (1983). *Clinical Trials: A Practical Approach.* Wiley: New York.

Powell, M.J.D. (1970). A Hybrid Method for Nonlinear Algebraic Equations. In: 'Numerical Methods for Nonlinear Algebraic Equations', Rabinowitz, P. (ed). Gordon and Breach, 1970.

Racine-Poon, A., Grieve, A.P., Fluhler, H. & Smith, A.F.M. (1986). Bayesian methods in practice: Experiences in the pharmaceutical industry (with Discussion). *Appl. Statist.*, **35**, 93-150.

Racine-Poon, A., Grieve, A.P., Fluhler, H. & Smith, A.F.M. (1987). A two-stage procedure for bioequivalence studies. *Biometrics*, **43**, 847-56.

Robins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* **58**, 527-35.

Schneiderman, M.A. & Armitage, P. (1962). A family of closed sequential procedures. *Biometrika*, **49**, 41-56.

Siegmund, D. (1979). Corrected diffusion approximations in certain random walk problems. *Advances in Applied Prob.* **11**, 701-19.

Sobel, M. & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Ann. Math. Statist.*, **20**, 502-22.

Spiegelhalter, D.J., Freedman, L.S. & Blackburn, P.R. (1986). Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials* **7**, 8-17.

Tsiatis, A.A., Rosner, G.L. & Mehta, C.R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40**, 797-803.

Wald, A. (1947). *Sequential Analysis*. New York: Wiley.

Wald, A. & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Statist.* **19**, 326-39.

Wang, S.K. & Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**, 193-9.

Weiss, L. (1962). On sequential tests which minimize the maximum expected sample size. *J. Amer. Statist. Assoc.* **57**, 551-66.

Whitehead, J. (1983). *The Design and Analysis of Sequential Clinical Trials.* Chichester: Horwood.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrics* **73**, 573-81.

Whitehead, J. & Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics* **39**, 227-36.